HANDBOOK OF ECONOMIC GROWTH

PHILIPPE AGHION AND STEVEN DURLAUF, EDITORS
ELSEVIER, 2005

Here is a list of chapters that are available online in working paper form. Thanks to Mark Feldman for reorganizing the list to match the order of the published handbook. Apologies for any errors or omissions; please let me know about them and I will correct them promptly.

Last updated March 14, 2006 by Chad Jones <chad@econ.berkeley.edu>.

The published version of the Handbook of Economic Growth is here. (Your university must have a subscription to the Handbooks for this to work.)

Ch 00 Reflections on Growth Theory (Not available online, to my knowledge.)
Robert M. Solow

Ch 01 Neoclassical Models of Endogenous Growth: The Effects of Fiscal Policy, Innovation and Fluctuations
Larry E. Jones and Rodolfo E. Manuelli

Ch 02 Growth with Quality-Improving Innovations: An Integrated Framework
Philippe Aghion, Harvard University and Peter Howitt, Brown University

Ch 03 Horizontal Innovation in the Theory of Growth and Development
Gino Gancia, CREI and Fabrizio Zilibotti, Institute for International Studies and University College London.

Ch 04 From Stagnation to Growth: Unified Growth Theory
Oded Galor, Hebrew University at Jerusalem and Brown University

Ch 05 Poverty Traps
Costas Azariadis, UCLA and John Stachurski, University of Melbourne

Ch 06 Institutions as the Fundamental Cause of Long-Run Growth
Daron Acemoglu, MIT, Simon Johnson, MIT, and James Robinson, UC Berkeley

Ch 07 Growth Theory through the Lens of Development Economics
Abhijit Banerjee, MIT and Esther Duflo, MIT

Ch 08 Growth Econometrics
Steven Durlauf, University of Wisconsin, Paul Johnson, Vassar College, and Jonathan Temple, University of Bristol

Ch 09 Accounting for Cross-Country Income Differences 
Francesco Caselli, Harvard University

Ch 10 Accounting for Growth in the Information Age
Dale Jorgenson, Harvard University

Ch 11 Externalities and Growth
Peter Klenow, Stanford University, and Andres Rodriguez-Clare, IADB

Ch 12 Finance and Growth: Theory and Evidence 
Ross Levine, University of Minnesota

Ch 13 Human Capital and Technology Diffusion
Jess Benhabib, NYU and Mark Spiegel, FRBSF

Ch 14 Growth Strategies
Dani Rodrik, Harvard University

Ch 15 National Policies and Economic Growth: A Reappraisal
William Easterly, NYU

Ch 16 Growth and Ideas 
Charles I. Jones, University of California at Berkeley

Ch 17 Long-term Economic Growth and the History of Technology
Joel Mokyr, Northwestern

Ch 18 General Purpose Technologies
Boyan Jovanovic, NYU and University of Chicago and Peter L. Rousseau, Vanderbilt

Ch 19 Technological Progress and Economic Transformation 
Jeremy Greenwood, University of Rochester and Ananth Seshadri, University of Wisconsin

Ch 20 The Effects of Technical Change on Labor Market Inequalities
Andreas Hornstein (Richmond Fed), Per Krusell (Princeton), and Gianluca Violante (NYU)

Ch 21 A Unified Theory of the Evolution of International Income Levels
Stephen Parente, University of Illinois and Edward Prescott, University of Minnesota

Ch 22 A Global View of Economic Growth

Jaume Ventura, CREI Pampeu Fabra

Ch 23 Trade, Growth and the Size of Countries
Alberto Alesina, Harvard University, Enrico Spolaore, Brown University, and Romain Wacziarg, Stanford University

Ch 24 Urbanization and Growth
J. Vernon Henderson, Brown University

Ch 25 Inequality, Technology, and the Social Contract
Roland Benabou, Princeton University

Ch 26 Social Capital
Steven N. Durlauf and Marcel Fafchamps

Ch 27 The Effect of Economic Growth on Social Structures
Francois Bourguignon, The World Bank

Ch 28 Economic Growth and the Environment: A Review of Theory and Empirics
William Brock, University of Wisconsin and M. Scott Taylor, University of Wisconsin

---

Back to my homepage.

Neoclassical Models of Endogenous Growth: The Effects of Fiscal Policy, Innovation and Fluctuations*

Larry E. Jones
University of Minnesota and
Federal Reserve Bank of Minneapolis

Rodolfo E. Manuelli
University of Wisconsin

August, 2004

1 Introduction

Despite its role as the centerpiece of modern growth theory, the Solow model is decidedly silent on some of its basic questions: Why is average growth in per capita income so much higher now than it was 200 years ago? Why is per capita income so much higher in the member countries of the OECD than in the less developed countries (LDC) of the world? The standard implementation of the Solow model really has no answers for these questions except, perhaps for differences, across time and across countries in the production possibility set. This is typically summarized by differences in Total Factor Productivity (TFP). The fundamental reasons for why TFP might be different in different countries, or in different time periods is left open for speculation. If these differences are supposed to be due to differences in innovations, it is not made clear why access to these innovations should be different, nor is it noted that these innovations themselves are economic decisions– they have costs and benefits, and are made by optimizing, private agents.

This basic weakness in the Solow model (and its followers) was the driving force behind the development of the class of endogenous growth models. This literature has been wide and varied, with the models developed ranging from perfectly competitive, convex models to ones featuring a range of types of market failures (e.g., increasing returns, external effects, imperfectly competitive behavior by firms, etc.). But, a common feature that has been emphasized throughout is knowledge, or human capital, and its production and dissemination. In come cases, this has been

*Draft of a chapter for the forthcoming Handbook of Economic Growth edited by S. Durlauf and P. Aghion.

directly treated in the modeling, in others it has been more tangential, an important consideration for quantitative development, but less so for qualitative work. That this focus is essential follows from the fact that the Solow model already accurately reflects the quantitative limits of using models with only physical capital. (That is, capital's share is determined by the data to put us in the Solow range, technologically.) Although they differ in their details, in the end, what this class of models points to as differences in development are differences in social institutions across time and countries. Thus, countries that have weaker systems of property rights, or higher wasteful taxation and spending policies, will tend to grow more slowly. Moreover, these differences in performance can be permanent if these institutions are unchanging. As a corollary, those countries who developed these growth enhancing institutions more recently (and some still have not), have levels of income that are lower than those in which they were adopted earlier, even if current growth rates show only small differences.

In this paper we limit ourselves to studying neoclassical models. By this we mean models with convex production sets, well behaved preferences and a market structure that is consistent with competitive behavior. Therefore, we do not review the large literature that adresses the role of externalities and non-competitive markets. As it turns out, most of the basic ideas behind this literature can be expressed in simple, convex models of aggregate variables without uncertainty. These are the models that are the first focus of this chapter. They have proven both highly flexible and easy to use. With them, we can give substance to statements like those above that property rights and other governmental institutions are key to long run growth rates in a society. Most of this branch of the literature is well known by now, and much of it appears on standard graduate macro reading lists. Accordingly, our discussion will be fairly brief.[1] One important, and as of yet unresolved issue, is the size of the growth effects of cross-country differences in fiscal policy. Thus, our review of the standard convex model is complemented with a discussion of the more recent findings about the quantitative effects of taxes (and government spending) on growth. Even though the theoretical effects of social institutions are well understood, this is less true of the recent work on perfectly competitive models of innovation, and so, comparatively more space is used to discuss that ongoing development. As a second focus, one issue that comes immediately to light in studying this class of models is the possibility that uncertainty per se might have an impact on long run performance. This points to the possibility that instability in property rights and institutions might change the incentives for investment. That is, how are the time paths of savings, consumption and investment affected by uncertainty in this class of models? How does this compare with how uncertainty affects decisions in the Solow model (i.e., Brock and Mirman

[1] Other authors have also presented comprehensive surveys of this literature (see Barro and Sala-i-Martin (1995), Jones and Manuelli (1997), and Aghion and Howitt (1998) for examples). Our aim is to complement those presentations, rather than repeat them, and hence, our focus is somewhat distinct.

(1972) vs. Cass (1965) and Koopmans (1965))?

Much less is known about the answers to these questions at the present time and that knowledge that does exist is much less widespread. For this reason, we present a fairly detailed discussion of the properties of stochastic, convex models of endogenous growth. To this end, we study models in which technologies and policies are subject to random shocks. We characterize the effects of differential amounts of uncertainty on average growth. We show that increased uncertainty can increase or decrease average growth depending on both the parameters of the model and the source of the uncertainty. A separate, but related topic, is the business cycle frequency properties of these models. This is left to future work.

In section 2, we lay out the basics of the class of neoclassical (i.e., convex) models of endogenous growth. We show how differences in social institutions across time and across countries can give rise to different performance, even over the very long run. We also lay out some of the interpretations of the model, including human capital investment and innovation and knowledge diffusion sectors, that lend richness to its interpretation. Section 3 discusses properties of the models when uncertainty is added, and shows how this can affect the long run growth rate of an economy.

2 Endogenous Growth: Infinite Lifetimes

Historically, the engine of growth as depicted in Solow's seminal work on the topic (1956) was the assumption of exogenous technical change. Thus, initially, growth models aimed at being consistent with growth facts, but gave up on the possibility of explaining them. Moreover, this approach has weaknesses in two distinct areas. First, it is difficult using the exogenous growth model to explain the observed long run differences in performance exhibited by different countries. Second, the productivity changes that are assumed exogenous in the Solow model are, in fact, the result of conscious decisions on the part of economic agents. If this is the case, it is then important to explore both the mechanism through which productivity changes as well as the factors that can give rise to the observed long run differences if we are to understand these phenomena. In this section we briefly review the basic optimal growth model as initially analyzed by Cass (1965) and Koopmans (1965). We then discuss the nature of the technologies consistent with endogenous growth and the role of fiscal policy in influencing the growth rate. We conclude with an analysis of the role of innovation in the context of convex models of equilibrium growth.

2.1 Growth and the Solow Model

In the simplest time invariant version of the Solow model, it can be shown that the per capita stock of capital converges to a unique value independent of initial conditions. It is then necessary to assume some exogenous source of productivity growth in order to account for long run growth. In Solow (1956), it is assumed that

labor productivity grows continually and exogenously. In response, the capital stock (assumed homogeneous over time) is continually increased allowing for a continual expansion in the level of output and consumption. The literature on endogenous growth has concentrated on replacing this assumed exogenous productivity growth by an endogenous process. If this change in productivity of labor is thought to arise from the invention of techniques consciously developed, the literature on endogenous growth can then be thought of as explicitly modeling the decisions to create this technological improvement (see Shell (1967) and (1973)). For this to go beyond a reinterpretation of the Solow treatment, it must be that the technology for discovering and developing these new technologies does not have itself a source of exogenous technological change. Because of this, these models all feature technologies that are time stationary.

The consumer problem in the simple growth model is given by

$$\max_{\{c_t\}} \sum_{t=0}^{\infty} \beta^t u(c_t)$$

subject to

$$\sum_{t=0}^{\infty} p_t(c_t + x_t) \leq W_0 + \sum_{t=0}^{\infty} p_t r_t k_t, \tag{1a}$$

$$k_{t+1} = (1 - \delta_k)k_t + x_t, \tag{1b}$$

where c_t is the level of consumption, x_t is investment, k_t is the capital stock, p_t is the price of consumption (relative to time 0), and r_t is the rental price of capital, all in period t, and W_0 is the present value of wealth net of capital income. The first order condition for (an interior) solution to this problem is just

$$u'(c_t) = \beta u'(c_{t+1})[1 - \delta_k + r_{t+1}]. \tag{2}$$

If, as is standard in the literature, the instantaneous utility function, $u(c_t)$, is assumed strictly concave, growth —defined as a situation in which $c_{t+1} > c_t$— requires

$$\beta[1 - \delta_k + r_{t+1}] > 1. \tag{3}$$

Condition (3) is fairly general, and must hold *independently of the details of the production side* of the economy. Thus, if the economy is going to display long run growth, the rate of return on savings must be sufficiently high.

What determines the economy's rate of return? In the standard Solow growth model —and in many convex models— firms can be viewed as solving a static problem. More precisely, each firm maximizes profits given by

$$\Pi_t = \max_{k,n} c + x - r_t k - w_t n,$$

subject to

$$c + x \leq F(k, n),$$

where F is a concave production function that displays constant returns to scale.

Since in equilibrium the household offers inelastically one unit of labor, the rental rate of capital must satisfy

$$r_t = f(k_t), \tag{4}$$

where $f(k) = F(k, 1)$, and k is capital per worker.

It is now straightforward to analyze growth in the Solow model. The equilibrium version of (2) is just

$$u'(c_t) = \beta u'(c_{t+1})[1 - \delta_k + f(k_{t+1})]. \tag{5}$$

If the *productivity of capital* is sufficiently low as the stock of capital per worker increases, then there is no long run growth. To see this, note that if $\lim_{k\to\infty} f'(k) = \underline{r}$, with $1 - \delta_k + \underline{r} < 1$, there exists a finite k^* such that $1 - \delta_k + f(k^*) = 1$. It is standard to show that the unique competitive equilibrium for this economy (as well as the symmetric optimal allocation) is such that the sequence of capital stocks $\{k_t\}$ converges to k^*. Given this, consumption is also bounded. (Actually, it converges to $f(k^*) - \delta_k k^*$.)

Can exogenous technological change 'solve' the problem. The answer depends on the nature of the questions that the model is designed to answer. If one is content to generate equilibrium growth, then the answer is a clear yes. If, on the other hand, the objective is to understand how policies and institutions affect growth, then the answer is negative.

To see this assume that technological progress is labor augmenting. Specifically, assume that, at time t, the amount of effective labor is $z_t = z(1 + \gamma)^t$. In order to guarantee existence of a balanced growth path we assume that the utility function is isoelastic (see Jones and Manuelli (1990) for details), and given by $u(c) = c^{1-\theta}/(1-\theta)$. Let a $\hat{}$ over a variable denote its value relative to effective labor. Thus, $\hat{c}_t \equiv c_t/(z(1+\gamma)^t)$. In this case, the balanced growth version of (2) is

$$(1 + \gamma)u'(\hat{c}_t) = \hat{\beta} u'(\hat{c}_{t+1})[1 - \delta_k + f'(\hat{k}_{t+1})]$$

where $\hat{\beta} = \beta(1 + \gamma)^{1-\theta}$.[2]

Standard arguments show that the equilibrium of this economy converges to a steady state $(\hat{c}, \hat{k})$. Thus, this implies that, asymptotically, consumption is given by $c_t = (1 + \gamma)^t z\hat{c}$. Thus, even though there is equilibrium growth, the growth rate is completely determined by the exogenous increase in labor augmenting productivity.

[2]Existence of a solution requires that $\beta(1 + \gamma)^{1-\theta} < 1$, which we assume.

2.2 A One Sector Model of Equilibrium Growth

As we argued before, the critical assumption that results in the economy not growing is that the marginal product of capital is low. The modern growth literature has emphasized the analysis of economies in which the marginal product of capital remains (sufficiently) bounded away from zero. This induces positive long-run growth in equilibrium. As we will show, how fast output grows in these models depends on a variety of factors (e.g., parameters of preferences). Because of this, these models have the property that the rate of growth is determined by the agents in the model.

Throughout, there will be one common theme. This mirrors the point emphasized above, that for growth to occur, the interest rate (either implicit in a planning problem or explicit in an equilibrium condition) must be kept from being driven too low. This follows immediately from the discussion above.

In terms of key features of the environment that are necessary to obtain endogenous growth there is one that stands out: it is necessary that the marginal product of *some* augmentable input be bounded strictly away from zero in the production of some augmentable input which can be used to produce consumption.

Since we are dealing with convex economies, the arguments in Debreu (1954) apply to the environments that we study. Thus, in the absence of distortionary government policies, equilibrium and optimal allocations coincide. Thus, for ease of exposition, we will limit ourselves to analyzing planner's problems.

The planner's problem in the basic one sector growth model is given by

$$\max_{\{c_t\}} \sum_{t=0}^{\infty} \beta^t u(c_t),$$

subject to

$$\begin{aligned} c_t + x_t &\leq F(k_t, n_t), \\ k_{t+1} &\leq (1-\delta_k)k_t + x_t, \end{aligned}$$

where c_t is per capita consumption, k_t is the per capita stock of capital, x_t is the (nonnegative) flow of investment, and n_t is employment at time t. Since we assume that leisure does not yields utility, the optimal (and equilibrium) level of n_t equals the endowment, which we normalize to 1. The Euler equation for this problem is just (5) given that. as before, we set $f(k) = F(k, 1)$. It follows that if $\lim_{k\to\infty} \beta[1-\delta_k+f(k)] > 1$, then $\limsup_t c_t = \infty$. Thus, there is equilibrium growth. This result does not depend on the assumption of just one capital stock. More precisely, in the case of multiple capital stocks, the feasibility constraint is just

$$\begin{aligned} c_t + \sum_{i=1}^{I} x_{it} &\leq f(k_{1t}, ..., k_{It}), \\ k_{it+1} &\leq (1-\delta_{ik})k_{it} + x_{it}. \end{aligned}$$

In this case, the natural analogue of the assumption that the marginal product of capital is bounded is just that there is a homogeneous of degree one function —a linear function— that is a lower bound for the actual production function. However, it turns out that all that is required is that there exist a ray that has bounded marginal products. Formally, this corresponds to

Condition 1 (G) *Assume that* $f(k_1, ..., k_I) \geq h(k_1, ..., k_I)$*, where* h *is concave, homogeneous of degree one and* C^1 *for all* $(k_1, ..., k_I) \in \mathbb{R}^I_+$*. Moreover, assume that there exists a vector* $\hat{k} = (\hat{k}_1, ..., \hat{k}_I)$*,* $\hat{k} \neq 0$*, such that if* $\hat{k}_i > 0$*,*

$$\beta[1 - \delta_k + h_i(\hat{k})] > 1, \qquad i = 1, ..., I$$

The basic result is the following (see Jones and Manuelli (1990))

Proposition 2 *Assume that Condition G is satisfied. Then, any optimal solution* $\{c_t^*\}$ *is such that* $\limsup_t c_t^* = \infty$.

As Jones and Manuelli (1990) show, the planner's solution can be supported as a competitive equilibrium. An extension to multiple goods is presented by Kaganovich (1998) and it is based on similar insights. It is clear that Condition G does not rule out decreasing returns to scale. This, in turn implies that this class of models is consistent with a version of the notion of conditional convergence: relatively poor countries are predicted to grow faster than richer countries, with the consequent closing of the income gap. Put it differently, theory suggests that, with a finite amount of data, it is difficult to distinguish an endogenous growth model from a Cass-Koopmans exogenous growth model. The main difference lies in the tail behavior of the relevant variables (output or consumption), and not in the balanced (or unbalanced) nature of the equilibrium path.

2.3 Fiscal Policy and Growth

In this section we describe the effects of taxes and government spending on the long run growth rate. Consider the problem faced by a representative agent

$$\max \sum_{t=0}^{\infty} \beta^t u(c_t, 1 - n_t)$$

subject to

$$(1+\tau^c)c_t + (1+\tau^x)p_t x_{kt} + (1+\tau^h)q_t \leq w_t(1-\tau^n)(n_{ct}h_t + n_{kt}h_t) + (1-\tau^k)r_t k_t + T_t + \Pi_t,$$

where τ^j represent tax rates, c_t is consumption, x_{kt} is investment in physical capital, q_t are market goods used in the production of human capital, $n_{it}h_t$ is effective labor

—the product of human capital and hours— allocated to sector i, k_t is the stock of capital, T_t is a government transfer, and Π_t are net profits.

Accumulation of human capital at the household level satisfies

$$h_{t+1} \leq (1-\delta_h)h_t + F^h(q_t, n_{ht}h_t),$$

where F^h is homogeneous of degree one, concave and increasing in each argument.

The economy has two sectors: producers of capital and consumption goods. Output of the capital goods industries satisfies

$$x_t \leq F^k(k_{kt}, n_{kt}h_t),$$

where F^k is homogeneous of degree one and concave.

Feasibility in the consumption goods industry is given by

$$c_t \leq F^c(k_{ct}, n_{ct}h_t),$$

where F^c is increasing and concave. It is not necessary to assume that this production function displays constant returns to scale.

It is illustrative to consider several special cases. Throughout, we assume that the utility function is of the form that is consistent with the existence of a balanced growth path. Specifically, we assume that $u(c,\ell) = (cv(\ell))^{1-\theta}/(1-\theta)$. Moreover, since our emphasis is on the role of taxes and tax-like wedges between marginal rates of substitution and transformation, we assume that lump sum transfers, T_t, are adjusted to satisfy the government budget constraint.

Case I: One Sector Model with Capital Taxation We assume that the consumer supplies one unit of labor inelastically. In this case $F^c = F^k = Ak + \hat{F}(k)$, where $\hat{F}(k)$ is strictly concave and $\lim_{k\to\infty} \hat{F}'(k) = 0$. For now we ignore human capital and set $F^h \equiv 0$. It follows that the balanced growth rate satisfies

$$\gamma^\theta = \beta[1 - \delta_k + \frac{1-\tau^k}{1+\tau^x}A].$$

Thus, in this setting, increases in the effective tax on capital, $(1-\tau^k)/(1+\tau^x)$ unambiguously decrease the equilibrium tax rate. Thus, unlike exogenous growth models, government policies affect the growth rate. Moreover, this simple example illustrates the size of the impact of changes in tax rates on the long run growth rate depend on the intertemporal elasticity of substitution $1/\theta$. More precisely the elasticity of the growth rate with respect to τ^k is given by

$$\frac{\partial\gamma}{\partial\tau^k}\frac{\tau^k}{\gamma} = -\frac{1}{\theta}\frac{\frac{\tau^k}{1+\tau^x}A}{1-\delta_k+\frac{1-\tau^k}{1+\tau^x}A}.$$

It follows that, other things constant, high values of the intertemporal elasticity of substitution result in large changes in predicted growth rates in response to changes in tax rates. Thus, even an example as simple as this one illustrates that the quantitative predictions of this class of models will heavily depend on the values of the relevant preference (and technology) parameters.

Case II: Physical and Human Capital: Identical Technologies In this section we assume that $F^c = F^k$, and $F^h = q$. This implies that all three goods —investment, consumption and human capital— are produced using the same technology and, in particular, the same physical to human capital ratio. As in the previous section, τ^k and τ^x do not play independent roles. Thus, to simplify notation, we will set $\tau^x = 0$. However, the reader should keep in mind that increases in the tax rate on capital income are equivalent to increases in the tax rate on purchases of capital goods.

In this case, the balanced growth conditions are

$$\gamma^\theta = \beta[1 - \delta_k + (1 - \tau^k)F_k(\kappa, n)] \tag{6a}$$

$$\frac{c}{h}\frac{v'(1-n)}{v(1-n)} = \frac{1-\tau^n}{1+\tau^c}F_n(\kappa, n) \tag{6b}$$

$$(1 - \tau^k)F_k(\kappa, n) - \delta_k = \frac{1-\tau^n}{1+\tau^h}F_n(\kappa, n)n - \delta_h \tag{6c}$$

$$\frac{c}{h} + (\gamma + \delta_k - 1) = F(\kappa, n). \tag{6d}$$

There are several interesting points. First, increases in the tax rate on consumption goods (i.e. sales or value added taxes) are equivalent to increases in the tax rate on labor income. Second, the relevant tax rate to evaluate the return on human capital is $(1 - \tau^n)/(1 + \tau^h)$. Thus, it is possible that increases in τ^n —as observed in the U.S. between the pre World War II and the post WWII periods— if matched by decreases in τ^h (corresponding, for example, to expansion in the quantity and quality of free public education) have no effect on the physical capital - human capital ratio, κ. Third, it is possible to show that increases in τ^k, τ^n, τ^h or τ^c result in lower growth rates. Last, without making additional assumptions about preferences and technology, it is not possible to sign the impact of changes in tax rates on other endogenous variables.

Case III: Physical and Human Capital: Different Factor Intensities In this case, we assume that only human capital is used in the production of human capital. Thus, $F^h = A_h n_h h$. This is the technology proposed by Uzawa (1964) and popularized in this class of models by Lucas (1988). For simplicity, we only consider capital and labor taxes. The relevant steady state conditions are (6a), (6b), and (6d). However, (6c) becomes

$$\gamma^\theta = \beta[1 - \delta_h + A_h n] \tag{7}$$

In this version of the model, changes in labor income taxes, reduce growth through their impact on hours worked (relative to leisure). However, if total work time is inelastically supplied, i.e. $v(\ell) \equiv 1$, the growth rate is pinned down by

$$\gamma^{\theta} = \beta[1 - \delta_h + A_h].$$

Thus, in this setting (which corresponds to Lucas (1988) model without the externality, and to Lucas (1990)), taxes have no effect on growth. Increases in the tax rate on capital income simply change physical capital - human capital ratio and they leave the after tax rate of return unchanged. The reason for this extreme for of neutrality is that even though taxes on labor income reduce the returns from education, they also reduce the cost of using time to accumulate human capital (the value of time decreases with increases in taxes), and the two changes are identical. Thus, the cost-benefit ratio of investing in education is independent of the tax code.

2.3.1 Quantitative Analysis of the Effects of Taxes

Since the development of endogenous growth theory there have been several studies of the implications of substituting lump-sum taxes for a variety of distortionary taxes. Jones, Manuelli and Rossi (1993), analyze the optimal choice of distortionary taxes in several models of endogenous growth. In the case that physical and human capital are produced using the same technology and labor supply is inelastic, they find that for parameterizations that make the predictions of the model consistent with observations from the U.S., the potential growth effects of drastically reducing (eliminating in most cases) all forms of distortionary taxation is quite high. For their preferred parameterization the increase in growth rates is about 3%. They study a version of the model in which $F^c = F^k \neq F^h$., and the functions F^k andF^h are both of the Cobb-Douglas variety, but differ in the average productivity of capital. Jones, Manuelli and Rossi estimate the capital share parameter to be equal 0.36 in the consumption sector, and to be somewhere in the 0.40-0.50 range in the human capital production sector.[3] They also allow labor supply to be elastic. Their findings suggest that switching to an optimal tax code result in increases in yearly growth rates of somewhere between 1.5% and 2.0% per year. These are substantial effects.

The third experiment that they consider involves the endogenous determination (by the planner) of the level of government consumption. In this case, they revert back to the one sector version of the model, and they explore not only the consequences of changing the intertemporal elasticity of substitution, but they allow for varying elasticity of substitution between capital and human capital. For their preferred characterization, they also find growth effects of about 2% per year. Moreover, as in the other experiments, the predictions are quite sensitive to the details of the model

[3]Jones, Manuelli and Rossi (1993) calibrate this share. Since they study the sensitivity of their results to changes in other parameters (e.g. the intertemporal elasticity of substitution), the market goods share is not constant across experiments.

—in particular, to the choice of the intertemporal elasticity of substitution, and the degree of substitutability between capital and human capital.

Stokey and Rebelo (1995) undertake a thorough review of several models that estimate the growth impact of tax reform. They argue that in the U.S. tax rates in the post WWII period are significantly higher than in the pre WWII era. This conclusion is based on the increase in the revenue from income taxes as a fraction of GDP in the early 1940s. To reconcile the models with this evidence, they conclude that the human capital share in the production of human capital must be large, and that this sector must be lightly taxed. This description is close to the Case III above and, as argued before, it results in no growth effects[4]. Thus, in agreement with Lucas (1990) —and using a very similar specification of the human capital production technology— they conclude that changes in tax rates cannot have large growth effects.

This conclusion depends on several assumptions. First, that the U.S. evidence shows an increase in the general level of taxes after WWII. Second, that even if there is a tax increase, the additional revenue is used to finance lump-sum transfers. Third, that the balanced growth path is a good description of the pre and post WWII economy.

Measuring changes in the relevant marginal tax rates is a difficult task. Barro and Sahasakul (1986) using tax records compute average marginal tax rates for the U.S. economy. Their estimates, consistent with the Stokey and Rebelo assumption, show an increase in the 1940s. Even though the evidence about changes in the tax rate consistently points to an increase, the implications of this result for the model are not obvious. Consider, first, the uses of tax revenue. If, for example, additional income tax revenues (at the local level) are used to finance local publicly provided goods (e.g. education), then Tiebout-like arguments suggest that the 'tax effect' of a tax increase is zero. In the U.S. a substantial increase in government spending corresponds to increases in expenditures on education and, hence, the possibility of individuals sorting themselves to buy the 'right' bundle of publicly provided private goods cannot be ignored. A second quantitatively important public spending program in the post WWII era is Social Security. To the extent that benefits are dependent on contributions, the statutory tax rate on labor income used to finance social security overstates the true tax rate.[5] In this case, tax payments purchase the right to an annuity whose value is dependent on the payment. Finally, in a model with multiple tax rates an increase in a single tax does not imply, necessarily, a decrease in the growth rate. For the U.S. the evidence on the time path of capital income taxes is mixed. In a recent study, Mulligan (2003) argues that the tax rate on capital income has steadily fallen in the last 50 years. Similarly, Prescott and McGrattan (2003) and (2004) find that a decrease in the tax rate on corporate income —one form of

[4]The results are continuous in the parameters. Thus, for market goods share close to zero, as Stokey and Rebelo prefer, the growth effects are small.

[5]In a pay-as-you-go system, even if the share of total payments that an individual receives is sensitive to his contributions, the same effect obtains.

capital income— is instrumental in explaining the increase in the value of corporate capital relative to GDP. Overall, we find that the quantitative evidence on the time path of the relevant tax rates to be difficult to ascertain. More work is needed, with an emphasis on matching model and data.

The next section considers the effects of endogenous government spending and transitional effects.

2.3.2 Productive Government Spending

A Simple Balanced Growth Result In this section we study a simple one sector model that provides a role for productive government spending. Our discussion follows the ideas in Barro (1990). Assume that firm $i's$ technology is given by

$$y_{it} \leq A k_{it}^{\alpha} h_{it}^{\eta} G_t^{1-\alpha-\eta},$$

where k_{it} and h_{it} are the amounts of physical and human capital used by the firm, and G_t is a measure of productive public goods that firms take as given. The government budget constraint is balanced in every period, and it satisfies

$$G_t = \tau^k r_t K_t + \tau^h w_t H_t,$$

where τ^k and τ^h are the tax rates on capital and income, and r_t and w_t are rental prices. For simplicity we assume that the instantaneous utility function is given by

$$u(c) = \frac{c^{1-\theta} - 1}{1-\theta}.$$

We also assume that the technologies to produce market goods and human capital are identical. In this case, it is immediate to show that the equilibrium is fully described by

$$\begin{aligned} \delta_h - \delta_k &= A^{1/(\alpha+\eta)}(\alpha\tau^k + \eta\tau^h])^{(1-\alpha-\eta)/(\alpha+\eta)}[\eta(1-\tau^h)\kappa^{\alpha/(\alpha+\eta)} - \alpha(1-\tau^k)\kappa^{-\eta/(\alpha+\eta)}], \\ \gamma^{\theta} &= \beta[1 - \delta_k + \alpha(1-\tau^k)A^{1/(\alpha+\eta)}(\alpha\tau^k + \eta\tau^h)^{(1-\alpha-\eta)/(\alpha+\eta)}\kappa^{-\eta/(\alpha+\eta)}], \end{aligned}$$

where κ is the physical capital - human capital ratio.

Some tedious algebra shows that the growth rate is not a monotonic function of the tax rates. In general, there is no growth when taxes are either too low (not enough public goods are provided) or too high (the private returns to capital accumulation are too low). For intermediate values of the tax rates, growth is positive (if A is sufficiently high). Thus, in general, increases in tax rates need not result in lower growth if they are accompanied by changes in government spending. Thus, a variant of the model with endogenous government spending (or endogenous taxation and optimally chosen government spending) has potential to reconcile positive growth effects associated with the removal of distortions with the U.S. evidence.

What does the U.S. evidence show? In the U.S. there is a substantial increase in the ratio of government spending to GDP in the post WWII period on the order of 15%. Even ignoring defense related expenditures, the size of the federal government relative to output is close to 5% in the pre WWII period, and it increases steadily in the post war to reach about 20% of income. Of course, not all forms of government spending are productive, but if the trend in the productive component follows the trend in overall spending, ignoring changes in government spending result in biased estimates of the effects of distortions.

The Barro model is silent about the reasons why the desired ratio of (productive) government spending to GDP would increase. For this, it is necessary to have a model of the collective decision making mechanisms which is clearly beyond the scope of this chapter.

Progressive Taxes and Transition Effects Our discussion of the assumptions that suffice for sustained growth clearly shows that homogeneity of degree one is not necessary. In both theoretical and applied work it is common to appeal to linearity in order to ignore transitional dynamics (see, Bond, Wang and Yip (1996)) and Ladron de Guevara, Ortigueira and Santos (1997)) for analysis of the dynamics of endogenous growth models). However, when taking the model to the data, the assumption that the economy is on the balanced growth path may not be appropriate.

In this section we describe the results of Li and Sarte (2001). The basic insight from their model that is relevant for our discussion is that in the presence of heterogeneity in individual preferences and nonlinearities in the tax code, shocks to the tax regime (they consider an increase in the degree of progressivity of the tax code) that ultimately result in a decrease in the growth rate can have basically no effects for several decades.

The basic model that they consider is one in which goods are produced according to the following technology

$$Y_t \leq AK_t^{\alpha} L_t^{1-\alpha} G_t^{1-\alpha},$$

where K_t is capital at time t, L_t is the flow of labor, and G_t is a measure of productive public goods. All individuals have isoelastic preferences given by $u(c) = (c^{1-\theta} - 1)/(1-\theta)$, but they differ in their discount factors, β_i. Li and Sarte assume that each type has mass $1/N$, where N is population. The tax code is nonlinear. Given aggregate income Y, and individual income y_i, the *tax rate* is given by a function $\tau(z)$, where z is the ratio of individual to average income. In this application, Lin and Sarte assume that

$$\tau(\frac{y_i}{Y}) = \zeta(\frac{y_i}{Y})^{\phi}.$$

Note that the case of proportional taxes —the case discussed so far— corresponds to $\phi = 0$. In this setting, higher values of ϕ are interpreted as corresponding to more progressive tax codes. Individual income is defined as the sum of capital and labor income. Government spending is financed with revenue from income taxes. Li and

Sarte show that the equilibrium is the solution to the following system of equations

$$\begin{aligned}
\gamma^{\theta} &= \beta_i\{1-\delta+[1-(1+\phi)\zeta(\frac{y_i}{Y})^{\phi}]\alpha A^{1/\alpha}(\frac{G}{Y})^{(1-\alpha)/\alpha}, \qquad i=1,2,...I,\\
\frac{G}{Y} &= \sum_{i=1}^{I}\zeta(\frac{y_i}{Y})^{1+\phi}\frac{1}{N},\\
1 &= \sum_{i=1}^{I}\frac{y_i}{Y}\frac{1}{N}.
\end{aligned}$$

In this model, changes in the progressivity of the tax code affect the rate of return —this is the standard effect— as well as the distribution of income. It is this last effect that generates the slow adjustment. It is possible to show that an increase in ϕ decreases long run growth, γ.

Li and Sarte explore the dynamic effects of a one time increase in ϕ that result in a decrease in the growth rate of 1.5%. On impact, output growth increases because since the distribution of income does not adjust immediately, government revenues increase and this, in turn, increase output. As the low discount factor individuals adjust their relative income (an increase in progressivity affects them more than proportionally), government revenue (and spending decrease. For parameter values that are designed to mimic the U.S. economy, Li and Sarte find that the half-life of the adjustment is over 40 years. Thus, any test for breaks in the growth rate as suggested by models in which convergence is immediate would conclude (incorrectly) that the tax increase has no effects on growth.

It is difficult to evaluate how appropriate the Li and Sarte model is to study the impact of tax reform in the U.S. economy. However, it casts doubt on the approach by Stokey and Rebelo which ignores transitional dynamics. Models that rely on changes in tax rates that, in turn, affect the distribution of income, are consistent with the view that the effects of those changes are not monotone, and that the full impact may not be felt for decades.

2.4 Innovation in the Neoclassical Model

One of the things that seems unsatisfactory to many economists in the presentation up to this point is the starkness with which the technological side of the model is described. As we argued above, the key in improving over the Solow model is to explicitly consider decisions made by private agents about investments they make that cause technology to improve. This both endogeneizes the growth process envisaged by Solow and breaks away from another key assumption of the exogenous growth literature, that technological change happens without any resource cost. But, much of the detail that one thinks about as being an important part of the innovation process seems to be missing from the simple convex models of growth described above. The idea that innovation is carried out by specialized researchers who pass

on their newfound knowledge to production line workers is just one example of this. Indeed, one question is whether or not that kind of structure is consistent with convex models of growth at all.

Because of this, in this section we describe a variant of the models presented in the last section that is more directed at identifying innovation as a special activity. The purpose of this exercise is not to fully exhaust the possibilities, but rather to show the reader that more is possible with the class of convex models than one might first think. In particular, since the model we will analyze is convex, standard price taking behavior is consistent with equilibrium behavior. In this sense, the example we will present is similar to the ideas developed by Boldrin and Levine (2002).

There are many models of innovation that do not have convex technology sets (e.g., see the surveys in Barro and Sala-i-Martin (1995) and Aghion and Howitt (1998)). In this setting, standard price taking behavior is either not consistent with equilibrium in those settings or they must include external effects. Because of this, all policy experiments in those models mix two conceptually distinct aspects of policy, the desire to correct for monopoly power and/or external effects and the distortionary effects of 'wedges' (e.g. taxes). This, in turn, implies that the answers to questions about the effects of alternative policies on both the incentives to innovate and overall welfare depends on the details of the specifications of external effects (e.g., do other researchers learn new innovations for free after one month, or one decade) and/or market power (e.g., is there only one researcher at the frontier and so a monopoly analysis is in order, or are there two, or many). Thus, one thing that a convex model of innovation has to add is answers to some of these questions which are less dependent on those details.

2.4.1 Notation

We will follow the notation above as closely as is possible. We assume that there are two types of labor supply available, researchers and workers. Each individual of each type has his own level of knowledge. We will assume that there is a continuum of identical households each with some researcher time and some worker time to supply to the market. These are given by L_1 researcher hours per household, and L_2 worker hours per household, where $L = L_1 + L_2$ is total labor supply within the household. We will assume that L_1 and L_2 are fixed, with no ability to move hours between them. (In this sense, it might be easier to think of the household as being made up of L_1 researchers and L_2 workers.)

Each household has the level of knowledge H_t that they can use with researcher hours during period t. Thus, if households are symmetric, H_t symbolizes the absolute frontier of what 'society' knows at date t. Similarly, the level of knowledge for the average worker hour at date t is denoted by h_t. This will represent the average knowledge of those workers that work in the final goods sector below.

Final consumption at date t is denoted by c_t. We abstract from physical capital

to simplify.

Production Functions We will assume that:

$$\begin{aligned} H_{t+1} &= H_t + A_H L_{1t}^H H_t, \\ h_{t+1} &= (1-\delta_h)h_t + A_h(L_{1t}^h H_t)^\alpha (L_{2t}^h h_t)^{1-\alpha}, \\ c_t &= A_c L_{2t}^c h_t, \\ L_{1t}^H + L_{1t}^h &= L_1, \\ L_{2t}^h + L_{2t}^c &= L_2. \end{aligned}$$

This formulation is equivalent to one in which quality adjusted labor of the form $Z = LH$, (resp. $Z = Lh$) is employed in each activity.

The idea here is that $I_H = A_H L_{1t}^H H_t$ is new research and development or innovation, increasing H corresponds to learning more at society's highest level. Note that we have assumed that there is no depreciation– the level of frontier knowledge does not go backward (positive depreciation is easily included). If no innovation is done, $L_{1t}^H = 0$ for all t, $H_{t+1} = H_t$ for all t, that is, the frontier is static. In this case, h_t would also be bounded, and hence the level of output would be bounded above. In this sense, innovation is necessary for growth to occur in this model.

Similarly, we think of the $I_h = A_h(L_{1t}^h H_t)^\alpha (L_{2t}^h h_t)^{1-\alpha}$ technology as Education and/or Worker Training. This is where family members at the frontier spend part of their time educating the workers from the family on the use of new techniques. The more time researchers spend in I_h, the less time they have to spend in I_H, and hence workers are better prepared and more productive, but the frontier moves out more slowly. Note that increasing $L_{2t}^h h_t$, holding $L_{1t}^h H_t$ constant increases total output of worker productive knowledge (new h) but lowers the average product of frontier knowledge workers in educating (bigger classes give more total new training, but less output per student). Symmetrically, the more time that production line workers spend learning new knowledge, the less time they have available for production of current consumption goods.

Preferences We assume that each researcher/worker supplies one unit of labor to the market inelastically and that each has preferences of the form:

$$U(c) = \sum_t \beta^t u(c_t),$$

where $u(c) = c^{1-\sigma}/(1-\sigma)$. This model, although different in detail, shares one common critical feature with those above: linearity in the reproducible factors.

2.4.2 Balanced Growth Properties of the Model

Like many Ak, style models, this one has the feature that it converges to a Balanced Growth Path. Indeed if the initial levels of relative human capitals (H_0/h_0) are right

the economy is on the BGP in every period. Standard techniques can be used to characterize this BGP. After some algebra, we find:

$$\gamma^{\sigma} = \beta\left[1 + A_H L_1\right]$$

This equation gives γ as a function of the basic parameters σ, β, A_H, and L_1. By construction, the comparative statics of growth rates with respect to the deep parameters of the model are identical to what one would find in an Ak model. The one difference is the inclusion of the endowment of researcher hours, L_1, note that γ is increasing in L_1. That is, if one country had a higher proportion of researchers in its population, output would grow faster.

Since γ does not depend on the other parameters of the model, it can be shown that the only way income taxes affect growth rates here is through their effect on the R&D sector. That is, if we have a linear income tax (at rate τ) either on income generated in all sectors, or on income generated only in the H sector the growth rate will fall to:

$$\gamma_{\tau}^{\sigma} = \beta[1 + (1-\tau)A_H L_1]$$

In particular, if income from the h and/or c sectors are taxed, but that from the H sector is not, there are no effects on growth. This is reminiscent of the Lucas (1988) model and the 2-sector model in Rebelo (1991).

The amount of time spent on R&D on the BGP is given by:

$$\begin{aligned} L_1^H &= \frac{\left[\left(\beta\left[1 + A_H^*\right]\right)^{1/\sigma} - 1\right]}{A_H^*} L_1 = \frac{\left[\left(\beta\left[1 + A_H L_1\right]\right)^{1/\sigma} - 1\right]}{A_H L_1} L_1 \\ &= \left[\left(\beta\left[1 + A_H L_1\right]\right)^{1/\sigma} - 1\right] / A_H \end{aligned}$$

Thus, if we compare two countries with different discount factors, but identical in other respects, the one with the higher β will devote a higher fraction of its researcher time to innovation and a lower percentage to teaching. This causes worker productivity in the consumption sector to be lower at first (and consumption as well), but growing faster and hence, eventually overtaking the low β country.

As a second point, note that increases in A_h do not change γ (and neither do changes in A_c). Thus, in this case, L_1^H is not affected and so the time series of H_t will be identical. This implies that h_t must be higher. Thus, wages of both researchers and workers will be higher. This is similar in spirit to the result in Boldrin and Levine (2002) that improvements in the copying technology raises the value of being an innovator. Since the only 'copying' being done here is the passing on of new knowledge to final goods workers, this is analogous in this setting.

These are simple comparative statics exercises which are meant only to show that much intuition about the process of innovation, and its comparative statics properties with respect to incentives can be illustrated in this class of models.

There are many interesting extensions of the analysis that one could imagine. These include heterogeneity among households (e.g., some researcher households, some worker households), the inclusion of uncertainty about the results of researcher time (and the questions that this raises about ex post hold up problems when one researcher is the 'first' discoverer), the training of researchers by other researchers when they have different H_t 's, the inclusion of more than one good or process, what types of Industrial Organization are possible through decentralizations of the allocation as a competitive equilibrium (e.g., firms specializing in R&D vs. each firm having an R&D division), etc.

But, the reader can see that much of the analysis will go through unchanged. Notable exceptions are when there are assumed to be external effects in the learning process. The simplest example of this here would be to assume that $h = H$ no matter what L_2^h is. In this case, unless this is completely internalized within a firm (i.e., there are no spillovers across firms) the Planner's problem will not be implementable as a competitive equilibrium.

2.4.3 Adding a Non-Convexity

Most models of innovation differ from the one outlined above in that they assume that there is a non-convexity in the innovation technology. There are two ways to include this in the specification above, and the differences between them highlight a key question about innovation.

These are:

$$H_{t+1} = H_t + A_H L_{1t}^H H_t if L_{1t}^H \geq L^*,\ H_{t+1} = H_t\ if\ L_{1t}^H < L^* \tag{8}$$

and

$$H_{t+1} = H_t + A_H (L_{1t}^H - L^*) H_t\ if\ L_{1t}^H \geq L^*,\ H_{t+1} = H_t\ if\ L_{1t}^H < L^*. \tag{9}$$

Although these two specifications look similar, they differ in one key aspect. The technology in (8) is convex anytime R&D is 'active' (i.e., $L_{1t}^H \geq L^*$). The technology in (9) has constant marginal costs when R&D is active, but features a set-up cost as well (given by L^* denominated in labor units). Technology (9) is the specification that is most commonly employed in the R&D literature while that in (8) is similar in spirit to that used in Boldrin and Levine. Because of this difference, all of the analysis outlined above can be used if the technology is that given in (8) so long as in the solution to the planner's problem we have that $NL_{1t}^H \geq L^*$ for all t where N is the number of households. That is, the allocation can be supported as a competitive equilibrium with price taking behavior, etc. There are some restrictions on the implicit Industrial Organization in the equilibrium however. For example, if $NL_{1t}^H = L^*$ for all t it follows that there can be at most one R&D firm in any equilibrium (or one firm with an R&D division). This, were it true, would cause serious concern for the price-taking assumption in the decentralization.

One interesting implication of this model is that whether or not the solution to the planning problem above (without the non-convexity) describes the competitive allocation depends on the size of the country, N. Thus, large countries would, in equilibrium, conduct R&D while small countries would not. Adding in a fourth sector in which researchers in large countries could train researchers in small countries would be a natural extension in which R&D was done in large countries, these researchers train high H workers from small countries (e.g., in Engineering schools), those newly trained 'researchers' return to their home countries where they subsequently train production line workers, etc.

This description of equilibrium cannot be true for (9), however. Price taking behavior in this setting implies that prices for the rental of new knowledge equal their marginal cost of production. This implies that there is no way to recover the set up cost of researchers spending L^* hours. Thus, there can be no competitive equilibrium. It follows that there must be some monopoly rent generated somewhere to decentralize any allocation. Typically this will be accompanied by inefficiencies and incorrect incentives to conduct R&D.

3 Fluctuations and Growth

3.1 Introduction

In this section we describe the existing results on the effects of 'volatility,' both in technologies and policies, on the long-run growth rate. We start with a brief summary of the empirical research in this area, and we then describe some simple theoretical models that are useful in understanding the empirical results. We end with the description of some recent work based on the theoretical models but aimed at evaluating their ability to *quantitatively* match the growth observations. As before, we ignore models based on aggregate non-convexities, and with non-competitive market structures.

3.2 Empirical Evidence

A relatively small (but growing) empirical literature has tried to shed light on the relationship between 'instability' and growth. This literature has concentrated on estimating reduced form models that try to capture, with varying degrees of sophistication, how 'volatility' (defined in a variety of different ways) affects long-run growth.

Kormendi and Meguire (1985) is probably the earliest study in this literature. They consider a sample of 47 countries with data covering the 1950-1977 period. Their methodology is to run a cross-country growth regression with the average (over the sample period) growth rate as the dependent variable, and a number of control variables, including the standard deviation of the growth rate (one measure of instability), as well as the standard deviations of policy variables such as the inflation

rate and the money supply. Kormendi and Meguire find that the coefficient of the volatility measure (the standard deviation of the growth rate) is *positive* and significant. Thus, a simple interpretation of their results is that more volatile countries —as measured by the standard deviation of their growth rates— grow at a higher rate.

Grier and Tullock (1989) use panel data techniques on a sample of 113 countries covering a period from 1951 to 1980. Their findings on the effect of volatility on growth are in line with those of Kormendi and Meguire. They find that the standard deviation of the growth rate is *positively,* and significantly, associated with mean growth rates.

Ramey and Ramey (1995) first report the results of regressing mean growth on its standard deviation on a sample of 92 countries as well as a subsample of 25 OECD countries, covering (approximately) the 1950-1985 period. They find that for the full sample the estimated effect of volatility is negative and significant, while for the OECD subsample the point estimate is positive, but insignificant. In order to allow for the variance of the innovations to the growth rate to be jointly estimated with the effects of volatility, Ramey and Ramey posit the following statistical model

$$\gamma_{it} = \beta X_{it} + \lambda \sigma_i + u_{it} \tag{10}$$

where X_{it} is a vector of variables that affect the growth rate and

$$u_{it} = \sigma_i \epsilon_{it}, \qquad \epsilon_{it} \sim N(0,1). \tag{11}$$

The model is estimated using maximum likelihood. The control variables used were the (average) investment share of GDP (Average I/Y), average population growth rate (Average γ_{Pop}), initial human capital (measured as secondary enrollment rate, H_0), and the initial level of per capita GDP (Y_0). They study separately the full sample (consisting of 92 countries) as well as a subsample of 25 OECD (more developed) economies. Their results are reproduced in columns (1) and (3) of Table I.

Table I: Growth and Volatility I

Variables	(1) (92−*Country*) N = 2,184	(2) (92−*Country*) N = 2,184	(3) (*OECD*) N = 888	(4) (*OECD*) N = 888
Constant	0.07 (3.72)	0.08 (3.73)	0.16 (5.73)	0.16 (4.48)
σ_i	−0.21 (−2.61)	−0.109 (−1.22)	−0.39 (−1.92)	−0.401 (−1.93)
Average I/Y	0.13 (7.63)	0.12 (6.99)	0.07 (2.76)	0.071 (2.67)
Average γ_{Pop}	−0.06 (−0.38)	−0.115 (−0.755)	0.21 (0.70)	0.230 (0.748)
H_0	0.0008 (1.18)	0.0007 (1.03)	0.0001 (2.00)	0.0001 (1.954)
Y_0	−0.009 (−3.61)	−0.009 −3.53)	−0.017 (−5.70)	−0.017 (−4.7445)
$\sigma_{\ln(I/Y)}$	-	−0.023 (1.81)		0.007 (0.22)

Note: t-statistics in parentheses
Source: Columns (1) and (3) Ramey and Ramey (1995)
Columns (2) and (4), Barlevy (2002)

For both sets of countries, Ramey and Ramey find that the standard deviation of the growth rate is *negatively* related to the average growth rate. However, for the OECD subsample, the coefficient is less precisely estimated (even though the point estimate is larger in absolute value). Ramey and Ramey also consider more 'flexible' specifications that try to capture differences across countries in the appropriate forecasting equations. Considering the most parsimonious version of their model, the estimated effect of volatility on growth is still positive. However, the strength of the estimated relationship is reversed: for the OECD subsample the point estimate is four times the size of the estimate for the full sample and highly significant.

In more recent work, Barlevy (2002) reestimates the Ramey and Ramey model with one change: he adds the standard deviation of the logarithm of the investment-output ratio ($\sigma_{\ln(I/Y)}$) as one of the explanatory variables. Barlevy hypothesizes that this variable can capture non-linearities in the investment function. His results, using the same basic data as Ramey and Ramey are in columns (2) and (4) of Table I.[6] For the full 92-country sample, the introduction of this measure of investment volatility halves the size of the coefficient of σ_i, and it is no longer significant at conventional levels. The coefficient on $\sigma_{\ln(I/Y)}$ is *negative* and significant (at 5%). For the OECD sample, the addition of $\sigma_{\ln(I/Y)}$ does not affect much the estimate of the effect of σ_i on growth. However, Barlevy points out that this is finding is not robust, since eliminating two outliers, Greece and Japan where high volatility of the investment share seems to be due to transitional dynamics, implies that neither the volatility of the growth rate nor $\sigma_{\ln(I/Y)}$ are significant.[7]

[6] We thank Gadi Barlevy for providing us the estimated coefficient for the control variables.
[7] The point estimates are negative but insignificant.

One possible explanation for the differences in the estimates of the effects of volatility on growth found in Kormendi and Meguire, Grier and Tullock and Ramey and Ramey, is —as pointed out by Ramey and Ramey and Barlevi— that Kormendi and Meguire and Grier and Tullock include among their explanatory variables the standard deviations of policy variables that could be proxying for $\sigma_{\ln(I/Y)}$.

Kroft and Lloyd-Ellis (2002) also start from the basic statistical model of Ramey and Ramey but offer a different way of decomposing volatility. They hypothesize that uncertainty can be split into two orthogonal components: uncertainty about changes in regime (e.g. expansion-contraction) and fluctuations within a given regime. To this end, they generalize the empirical specification of the Ramey and Ramey statistical model to account for this. They assume that

$$\gamma_{ist} = \beta X_{it} + \lambda_w \sigma_{iw} + \lambda_b \sigma_{ib} + \upsilon_{ist}, \tag{12a}$$

$$\upsilon_{ist} = \sigma_{iw}\epsilon_{it} + \mu_{is}, \qquad \epsilon_{it} \sim N(0,1), \tag{12b}$$

$$\mu_{is} = \begin{cases} \mu_{ie} & \text{with probability } p_i = \frac{T_{ie}}{T} \\ \mu_{ir} & \text{with probability } 1 - p_i \end{cases} \tag{12c}$$

Kroft and Lloyd-Ellis interpret the standard deviation of the random variable μ_{is}, σ_{ib} —which they assumed observed by the economic agents but unobserved by the econometrician— as a measure of variability *between* regimes, while σ_{iw} is viewed as the *within-regime* variability. Kroft and Lloyd-Ellis estimate their model by maximum likelihood using the same sample as Ramey and Ramey. The results are in Table II

Table II: Growth and Volatility II

Independent Variable	92-Country Sample (2,208 observations)	OECD Sample (888 observations)
Constant	0.00132 (0.022)	0.095 (1.89)
Within-phase volatility (σ_{iw})	2.63 (4.69)	0.90 (1.44)
Between-phase volatility (σ_{ib})	−2.65 (−6.35)	−1.11 (−2.33)
Average investment share of GDP	−0.01 (−0.26)	−0.004 (−0.073)
Average population growth rate	0.58 (1.24)	0.28 (0.62)
Initial human capital	0.001 (0.66)	−0.00001 (−0.096)
Initial per capita GDP	0.002 (0.25)	−0.0008 (−1.30)

Note: t-statistics in parentheses.
Source: Kroft and Lloyd-Ellis (2002).

The major finding is that the 'source' of volatility matters: Increases in σ_{iw} — the within phase standard deviation— have a positive impact on growth for the full sample. For the OECD, the coefficient estimate is still positive but about one third of the size. The effect of the between-phase volatility, σ_{ib}, is negative in both cases.

However, the effects are stronger for the full sample. It is not easy to interpret the phases identified by Kroft and Lloyd-Ellis in terms of a switching model because their estimation procedure assumes that the econometrician can identify whether a particular period corresponds to either a recession or an expansion.[8]. Kroft and Lloyd-Ellis also use the same controls as Ramey and Ramey. However, they find that, when the two variances are allowed to differ, none of the control variables is significant. It is not clear why this is the case. One possibility is that the 'phases' that they identify are correlated with the control variables (this seems like a likely situation in the case of investment). Another possibility is that the control variables, in the Ramey and Ramey formulation, capture the non-linearity associated with the regime shift and that, once the shifts are taken into account, the control variables have no explanatory power. In any case, this illustrates a point that we will come back to: the fragility of the "growth" regressions suggest that better theoretical models are necessary to more provide restrictions that will allow to identify the parameters of interest.

The results of Ramey and Ramey and Kroft and Lloyd-Ellis are consistent with the existence of nonlinearities in the relationship between measures of instability and growth. Fatás (2001) estimates a number of different specifications of the relationship between instability and growth. His approach is to run standard cross country regressions. His data set is taken from the most recent version of the Heston-Summers sample and includes 98 countries with information covering the period 1950-1998. His estimates (see Table III) support the view that the effect of volatility on growth is nonlinear. Using Fatás' basic estimate —shown in column (1) of Table III— the pure effect of volatility is *negative* with a coefficient of -2.772 indicating that a one standard deviation increase in volatility reduces the growth rate by over 2.5%. However, the interaction term, corresponding to the variable Volatility * GDP is positive and equal to 0.340. According to these estimates, the net effect of σ_i on γ_i for the *richest* countries in the data is *positive* and greater than 0.3. For the less developed countries the estimate of the effect of volatility is *negative*. Columns (2) and (3) use other measures of non-linearity (initial per capita GDP and M3/Y, a measure of financial development), with similar outcomes: In all cases there is a significant effect, and increases in volatility are less detrimental to growth —and could even have a positive effect— the more developed a country is according to the proxy variables.

[8]Kraft and Lloyd-Ellis estimate the probabilites p_i as the fraction of the time that an economy spends in the recession "phase," defined as periods of negative output growth. Thus, not only is the process assumed to be *i.i.d.* but the transition probabilites are not jointly estimated with the parameters.

Table III: Growth and Volatility III			
Independent Variable	(1)	(2)	(3)
Volatility (σ_i)	−2.772	−1.700	−0.270
	(0.282)	(0.645)	(0.091)
GDP per capita (1960)	−2.229	−1.856	−0.953
	(0.235)	(0.422)	(0.220)
Human capital (1960)	0.037	0.040	0.026
	(0.015)	(0.018)	(0.017)
Average investment share of GDP	0.083	0.143	0.120
	(0.013)	(0.021)	(0.024)
Average population growth rate	−0.624	−0.562	−0.465
	(0.153)	(0.205)	(0.465)
Volatility * GDP	0.340	−	-
	(0.036)		
Volatility * GDP (1960)	−	0.212	-
		(0.082)	
Volatility * M3/Y	-	-	0.004
			(0.001)
R^2	0.77	0.58	0.57

Note: Sample 1950-1998. Robust standard errors in parentheses
Source: Fatás (2001

Martin and Rogers (2000) also study the relationship between the standard deviation of the growth rate and its mean, in a cross section of countries and regions. They study two samples —European regions and industrialized countries— and in both cases they find a *negative* relationship between σ_γ and γ. However, when they consider a sample of developing countries the point estimates are positive, but in general insignificant.

It is not easy to explain the differences between Ramey and Ramey, Fatás and Martin and Rogers. The period used to compute the growth rates (1962-1985 for Ramey and Ramey, 1950-1998 for Fatás and 1960 to 1988 for Martin and Rogers), and the set of less developed countries included (68 in Ramey and Ramey's study, and 72 in Martin and Rogers') are fairly similar. The two studies differ on their definition of the growth rate (simple averages in the Ramey and Ramey and Fatás papers, and estimated exponential trend in Martin and Rogers), and in the variables that are used as controls. However, it is somewhat disturbing that what appear, in the absence of a theory, as ex-ante minor differences in definitions can result in substantial differences in the estimates.

Siegler (2001) studies the connection between volatility in inflation and growth rates and mean growth for the pre 1929 period. Specifically, he uses panel data methods for a sample of 12 (presently developed) countries over the 1870-1929 period. He finds that volatility and growth are negatively correlated, and this finding is robust to the inclusion of standard growth regression type of controls.

Dawson and Stephenson (1997) estimate a model similar to (10) and (11) applied to U.S. states. They use the average (over the 1970-1988 period) growth rate of gross state product per worker for U.S. states as their growth variable, and its standard deviation as a measure of volatility. In addition, they include in their cross-sectional

regression the standard (in growth regressions) control variables (investment rate, initial level of gross state product per worker, labor force growth rate, and initial human capital). Dawson and Stephenson find that volatility has *no impact* on the growth rate, once the other effects are included. Unfortunately, they do not report the 'raw' correlation between mean growth and its standard deviation. Thus, it is not possible to determine if the lack of significance is due to the use of controls, or is a more robust feature of U.S. states growth performance.

Mendoza (1997) differs from the previous studies in terms of his definition of instability. Instead of the standard deviation of the growth rate, which, in general, is endogenous, he identifies instability with the standard deviation of a country's terms of trade. He estimates a linear model using a cross section of countries and finds a *negative* relationship between instability and growth. His sample is limited to only 40 developed and developing countries, and it only covers the period 1971-1991.

A fair summary of the existing results is that there is no sharp characterization of the relationship between fluctuations and growth. Variation across studies in samples or specifications yield fairly different results. Moreover, the findings do not seem robust to details of how the statistical model is specified.

Are the empirical findings of the channel through which uncertainty affects growth more robust? Unfortunately, the answer is negative. Ramey and Ramey find that volatility —measured as the standard deviation of the growth rate— does not affect the investment-output ratio. More recently, Aizenman and Marion (1999) find that volatility is negatively correlated with investment, when investment is disaggregated between public and private. Fatás estimates a non-linear model of the effect of volatility on investment. He finds that increases in volatility decrease investment in poor countries, but that the opposite is true in high income countries. Thus his findings are consistent with the view that changes in volatility affect mean growth rates through (at least partially) their impact upon investment decisions.

How should these empirical results be interpreted? Even though it is tempting to take one's preferred point estimate as a measure of the impact of fluctuations (or business cycles) on growth there are two problems with this approach. First, the empirical estimates are not robust to the choice of specification of the reduced form. Second, and more important in our opinion, is that from the point of view of policy design, the relevant measures of volatility is the —in general unobserved— volatility in *policies and technologies*. In most models, the growth rate (and its standard deviation) are endogenous variables and, as usual, the point estimate of one endogenous variable on another is at best difficult to interpret. One way of contributing to the interpretation of the empirical results is to study what simple theoretical models predict for the estimated relationships. In the next section we present a number of very simple models to illustrate the possible effects of volatility in fundamentals on mean growth. In the process, we find that it is very difficult to interpret the empirical findings. To put it simply, there are theoretical models that —depending on the sample— do not restrict the sign of the estimated coefficient of

the standard deviation of the growth rate on its mean. Moreover, the sign and the magnitude of the coefficient is completely uninformative to determine the effect of volatility on growth.

3.3 Theoretical Models

The analysis of the effect of uncertainty on growth can be traced to the early work of Phelps (1962), and Levhari and Srinivasan (1969) who studied versions of the stochastic consumption-saving problem that are similar to the linear technology versions of endogenous growth models. More recently, Leland (1974), studies a stochastic Ak model, and he shows that the impact of increased uncertainty on the consumption/output ratio depends on the size of the coefficient of risk aversion.

Even in deterministic versions of models that allow for the possibility of endogenous growth, existence of equilibria (and even optimal allocations) requires strong assumptions on the fundamentals of the economy (see Jones and Manuelli (1990) for a discussion). At this point, there are no general results on existence of equilibrium in stochastic versions of those models. In special cases, most authors provide conditions under which an equilibrium exists (see Levhari and Srinivasan (1969), Mendoza (1997), Jones et. al (2003a) and (2003b) for various versions). A recent, more general result is contained in de Hek and Roy (2001). These authors consider fairly general utility and production functions, but limit themselves to $i.i.d.$ shocks. It is clear that more work is needed.

In what follows, we will describe a general linear model and we will use it to illustrate the predictions of the theory for the relationship between mean growth rates and their variability. To simplify the presentation we switch to a continuous time setting. In order to obtain closed-form solutions we specialize the model in terms of specifying preferences and technology. Moreover, we will limit ourselves to $i.i.d.$ shocks. Generalizations of these assumptions are discussed in the section that presents quantitative results.

3.4 A Simple Linear Endogenous Growth Model

We begin by presenting a stochastic analog of a standard Ak model with a 'twist.' Specifically, we consider the case in which there are multiple linear technologies, all producing the same good. In order to obtain closed-form results we specify that the utility of the representative household is given by

$$U = E\left[\int_0^\infty e^{-\rho t}\frac{c_t^{1-\theta}}{1-\theta}dt\,|F_0\right]. \tag{13}$$

We assume that each economy has two types of technologies to produce consumption (alternatively, the model can be interpreted as a two sector model with goods

that are perfect substitutes). Output for each technology satisfies

$$\begin{aligned} dk_t &= ((A-\delta_k)k_t - c_{1t})dt + \sigma_k k_t dW_t + \eta_k k_t dZ_t^k, && \text{(14a)} \\ db_t &= ((r-\delta_b)k_t - c_{2t})dt + \sigma_b b_t dW_t + \eta_b b_t dZ_t^b, && \text{(14b)} \end{aligned}$$

where (W_t, Z_t^k, Z_t^b) is a vector of three independent standard Brownian motion processes, and k_t and b_t are two different stocks of capital. This specification assumes that each sector is subject to an aggregate shock, W_t, as well as sector (or technology) specific shocks, Z_t^j.

To simplify the algebra, we assume that capital can be costlessly reallocated across technologies, and we denote total capital by $x_t \equiv k_t + b_t$. Setting (without loss of generality) $k_t = \alpha_t x_t$ (and, consequently $b_t = (1-\alpha_t)x_t$) it follows that total capital evolves according to

$$\begin{aligned} dx_t &= [(\alpha_t(A-\delta_k) + (1-\alpha_t)(r-\delta_b))x_t - c_t]dt + \\ &\quad [(\alpha_t\sigma_k + (1-\alpha_t)\sigma_b)dW_t + \alpha_t\eta_k dZ_t^k + (1-\alpha_t)\eta_b dZ_t^b]x_t. \end{aligned} \qquad \text{(15)}$$

Given the equivalence between equilibrium and optimal allocations in this convex economy, we study the solution to the problem faced by a planner who maximizes the utility of the representative agent subject to the feasibility constraint. Formally, the planner solves

$$\max E\left[\int_0^\infty e^{-\rho t}\frac{c_t^{1-\theta}}{1-\theta}dt\,|F_0\right],$$

subject to (15).

Let the value of this problem be $V(x)$. Then, it is standard to show that the solution to the planner's problem satisfies the Hamilton-Jacobi-Bellman equation

$$\rho V(x) = \max_{c,\alpha}\left[\frac{c^{1-\theta}}{1-\theta} + V'(x)(\mu(\alpha)x - c) + \frac{V''(x)x^2}{2}\sigma^2(\alpha)\right],$$

where

$$\begin{aligned} \mu(\alpha) &= r + \alpha(A-r) - (\alpha\delta_k + (1-\alpha)\delta_b), && \text{(16a)} \\ \sigma^2(\alpha) &= (\alpha\sigma_k + (1-\alpha)\sigma_b)^2 + \alpha^2\eta_k^2 + (1-\alpha)^2\eta_b^2. && \text{(16b)} \end{aligned}$$

It can be verified that the solution is given by $V(x) = v\frac{x^{1-\theta}}{1-\theta}$, where

$$v = \left[\frac{\rho - (1-\theta)[\mu(\alpha^*) - \delta(\alpha^*) - \theta\frac{\sigma^2(\alpha^*)}{2}]}{\theta}\right]^{-\theta} \qquad \text{(17)}$$

and $\delta(\alpha) = \alpha\delta_k + (1-\alpha)\delta_b$.

The optimal decision rules are

$$\alpha^* = \frac{\frac{A-\delta_k-(r-\delta_b)}{\theta} - \sigma_b(\sigma_k - \sigma_b) + \eta_b^2}{(\sigma_k - \sigma_b)^2 + \eta_b^2 + \eta_k^2}, \tag{18a}$$

$$c = \frac{\rho - (1-\theta)[\mu(\alpha^*) - \delta(\alpha^*) - \theta\frac{\sigma^2(\alpha^*)}{2}]}{\theta} x. \tag{18b}$$

For the solution to be well defined it is necessary that $\rho - (1-\theta)[\mu(\alpha^*) - \delta(\alpha^*) - \theta\frac{\sigma^2(\alpha^*)}{2}] > 0$, which we assume. (In each case we make enough assumptions to guarantee that this holds.)[9]

It follows that the equilibrium stochastic differential equation satisfied by aggregate wealth is given by

$$\begin{aligned} dx_t = & [\frac{\mu(\alpha^*) - (\delta(\alpha^*) + \rho)}{\theta} - (1-\theta)\frac{\sigma^2(\alpha^*)}{2}]x_t dt + \\ & [(\alpha^*(\sigma_k - \sigma_b) + \sigma_b)dW_t + \alpha_k^* \eta_k dZ_t^k + (1-\alpha^*)\eta_b dZ_t^b]x_t, \end{aligned} \tag{19}$$

and the instantaneous growth rate of the economy, γ, and its variance, σ_γ^2, satisfy

$$\gamma = \frac{\mu(\alpha^*) - (\delta(\alpha^*) + \rho)}{\theta} - (1-\theta)\frac{\sigma^2(\alpha^*)}{2}, \tag{20a}$$

$$\sigma_\gamma^2 = (\alpha^*(\sigma_k - \sigma_b) + \sigma_b)^2 + \alpha^{*2}\eta_k^2 + (1-\alpha^*)^2\eta_b^2. \tag{20b}$$

One is tempted to interpret (19) as the theoretical analog of (10) by defining the stochastic growth rate as

$$\gamma_t = \frac{dx_t}{x_t}.$$

Given this definition, the discrete time —with period length equal to one— version of the stochastic process followed by the growth rate is

$$\gamma_t = \frac{\mu(\alpha^*) - (\delta(\alpha^*) + \rho)}{\theta} - (1-\theta)\frac{\sigma_\gamma^2}{2} + \varepsilon_t, \tag{21a}$$

$$\varepsilon_t = (\alpha^*(\sigma_k - \sigma_b) + \sigma_b)dW_t + \alpha^*\eta_k dZ_t^k + (1-\alpha^*)\eta_b dZ_t^b. \tag{21b}$$

This simple model driven by i.i.d. shocks has a stark implication: the growth rate is i.i.d. and is independent of other endogenous (or exogenous) variables, except through the joint dependence on the error term. Using panel data, it is relatively easy

[9] In endogenous growth models existence of an equilibrium is not always guaranteed. The main problem is that with unbounded instantaneous utility and production sets, utility can be infinite. For a discussion of some conditions that guaranteee existence see Jones and Manuelli (1990) and Alvarez and Stokey (1998). The key issue is that the return function is unbounded above when $0 < \theta < 1$, and unbounded below if $\theta > 1$. In this setting, it can be shown that $c > 0$ is equivalent to ensuring boundedness.

to reject this implication. This, however, is not an intrinsic weakness of this class of models. The theoretical setting *can* be generalized to include serially correlated shocks and a non-linear structure, which could account for "convergence" effects, and would provide a role for lagged dependent variables. However, generalizing the theoretical model comes at the cost of not being able to discuss the impact of different factors on the growth rate, except numerically.

What is the (simple) class of model that we study useful for? We view the class of theoretical models that we present as more appropriate to discuss the implications of the theory for cross section regressions since, in this case, the constant $\frac{\mu(\alpha^*)-(\delta(\alpha^*)+\rho)}{\theta} - (1-\theta)\frac{\sigma_\gamma^2}{2}$ can be correlated with other variables like the investment-output ratio.

Even though there is a formal similarity between (21) and (10)-(11), the theoretical model suggests that the simple approach that ignores that the same factors that affect σ_γ, also influence the true value of β in (10) can result in incorrect inference. Alternatively, the "deep parameters" are not the means and the standard deviation of the growth rates. They are the means and standard deviations of the driving stochastic processes. In terms of those parameters, the "true" model is non-linear.

Whether the model in (21) implies a positive or negative relationship between fluctuations and growth depends on the sources of shocks. At this general level it is difficult to illustrate this point, but we will come back to it in the context of specific examples.

It is not obvious how to define the investment ratio in this model. The change in cumulative investment in k, X_k, is given by,

$$dX_{kt} = \delta_k k_t dt + dk_t,$$

while the change in total output can be defined as[10]

$$dY_t = \mu(\alpha^*)x_t dt + \sigma_\gamma dM_t,$$

where M_t is a standard Brownian motion defined so that

$$\sigma_\gamma dM_t = (\alpha^*(\sigma_k - \sigma_b) + \sigma_b)dW_t + \alpha^*\eta_k dZ_t^k + (1-\alpha^*)\eta_b dZ_t^b.$$

In order to avoid technical problems, we consider a discrete time approximation in which the capital stocks change only at the beginning of the period. The investment-output ratio (for physical capital) is given by

$$z_t = \frac{\gamma + \delta_k + \sigma_\gamma \varepsilon_t}{\mu(\alpha^*) + \sigma_\gamma \varepsilon_t},$$

[10] This is not the only possible way of defining output. It assumes that the economy two sectors (or technologies). However, another interpretation of this basic framework considers b_t as bonds, and k_t as the only real stock of capital. We will be precise about the notion of output in each application.

where ε_t is the same noise that appears in (21). Since the previous expression is non-linear, we approximate it by a second order Taylor expansion to obtain

$$z_t = \frac{\gamma + \delta_k}{\mu(\alpha^*)} + \frac{\sigma_\gamma[\mu(\alpha^*) - (\gamma + \delta_k)]}{\mu(\alpha^*)^2}\varepsilon_t - \frac{\sigma_\gamma^2[\mu(\alpha^*) - (\gamma + \delta)]}{\mu(\alpha^*)^3}\varepsilon_t^2. \tag{22}$$

The mean investment ratio, which we denote z, is given by

$$z = \frac{\gamma + \delta_k}{\mu(\alpha^*)}[1 + \frac{\sigma_\gamma^2}{\mu(\alpha^*)^2}] - \frac{\sigma_\gamma^2}{\mu(\alpha^*)^2}. \tag{23}$$

Given this approximation, the model implies that the covariance between the growth rate and the investment-output ratio is

$$cov(\gamma_t, z_t) = \frac{\sigma_\gamma^2[\mu(\alpha^*) - (\gamma + \delta_k)]}{\mu(\alpha^*)^2}, \tag{24}$$

while the standard deviation of z_t is

$$\sigma_z = \frac{\sigma_\gamma[\mu(\alpha^*) - (\gamma + \delta_k)]}{\mu(\alpha^*)^2}\frac{(1 + \mu(\alpha^*)^2)^{1/2}}{\mu(\alpha^*)}. \tag{25}$$

Simple algebra shows that, given that the existence condition (17) is satisfied, $cov(\gamma_t, z_t) > 0$. Thus, in a simple regression, the investment ratio has to appear to affect positively growth. At this general level it is more difficult to determine if high σ_z economies are also high γ economies. The problem is that there are a number of factors that jointly affect γ and σ_z. In order to be more precise, it is necessary to be specific about the sources of heterogeneity across countries. We will be able to discuss the sign of this relationship in specific contexts.

We now use this 'general' model to discuss —in a variety of special cases— the connection between the variability of the growth rate of output and its mean

3.4.1 Case 1: An Ak Model

Probably the simplest model to illustrate the role played by differences in the variability of the exogenous shocks across countries is the simple Ak model. Even though it is a special case of the model described in the previous section, it is useful to describe the technology in a slightly different way. Let the feasibility constraint for this economy be given by

$$\int_0^t \hat{A}k_s ds + \int_0^t \sigma_y \hat{A}k_s dW_s \geq \int_0^t c_s ds + \int_0^t dX_{ks}.$$

The left hand side of this condition is the accumulated flow of output until time t, and the right hand side is the accumulated uses of output, consumption and investment. The law of motion of capital is

$$dk_t = -\delta_k k_t dt + dX_{kt},$$

where δ_k is the depreciation rate. Expressing the economy's feasibility constraint in flow form, and substituting in the law of motion for physical capital, the resource constraint satisfies

$$dk_t = [(\hat{A} - \delta_k)k_t - c_t]dt + \sigma_y \hat{A} k_t dW_t. \tag{26}$$

The planner's problem —which coincides with the competitive equilibrium in this economy— is to maximize (13) subject to (26). This problem resembles the more general model we introduced in the previous section if we set $\eta_b = \sigma_b = \eta_k = 0$, and

$$\begin{aligned} A &= \hat{A} - \delta_k, \\ \sigma_k &= \sigma_y \hat{A}. \end{aligned}$$

In addition, we need to make sure that the "b" technology is not used in equilibrium. A simple way of guaranteeing this is to view $r - \delta_b$ as endogenous, and to choose it so that, in equilibrium, $\alpha^* = 1$; that is, all of the investment is in physical capital. It is immediate to verify that this requires

$$r - \delta_b = \hat{A} - \delta_k - \theta \sigma_y^2 \hat{A}^2.$$

In this case is it follows that $x_t = k_t$ and the formulas in (20) imply that the mean growth rate and the variance of the growth rate satisfy

$$\begin{aligned} \gamma &= \frac{\hat{A} - (\rho + \delta_k)}{\theta} - (1 - \theta)\frac{\sigma_\gamma^2}{2}, \\ \sigma_\gamma^2 &= \sigma_y^2 \hat{A}^2. \end{aligned}$$

This result, first derived by Phelps (1962) and Levhari and Srinivasan (1969), shows that, in general, the sign of the relationship between the variance of the technology shocks, σ_y^2, and the growth rate is ambiguous:

- If preferences display less curvature than the logarithmic utility function, i.e. $0 < \theta < 1$, increases in σ_y are associated with decreases in the mean growth rate, γ.
- If $\theta > 1$, increases in σ_y are associated with increases in the mean growth rate, γ.
- In the case in which the utility function is the log (this corresponds to $\theta = 1$) there is no connection between fluctuations and growth.

The basic reason for the ambiguity of the theoretical result is that the total effect of a change in the variance of the exogenous shocks on the saving rate —and ultimately on the growth rate— can be decomposed in two effects that work in different directions:

- An increase in the variance of the technology makes acquiring future consumption less desirable, as the only way to purchase this good is to invest. Thus, an increase in variance of the technology shocks has a *substitution effect* that increases the demand for current (relative to future) consumption. This translates into a lower saving and growth rates.

- On the other hand, an increase in the variability of the exogenous shocks induces also an *income effect*. Intuitively, for concave utility functions, the fluctuations of the marginal utility decrease with the level of consumption. Thus, the (negative) effect of fluctuations is smaller when consumption is high. This income effect increases savings, as this is the only way to have a 'high' level of consumption (i.e. to spend more time on the relatively flat region of the marginal utility function).

The formula we derived shows that the relative strength of the substitution and income effects depends on the degree of curvature of the utility function: if preferences have less curvature than the logarithmic function, the substitution effect dominates and increases in the variance of the exogenous shocks reduce growth. If the utility of the representative agent displays more curvature than the logarithmic function, the income effect dominates and the relationship between fluctuations and growth is positive.

In this simple economy, the variance of the technology shock, σ_y^2, and the variance of the growth rate of output, σ_γ^2, coincide up to scale factor $\hat{A}$.[11] If one views the differences across countries as due to differences in σ_y^2,[12] the theoretical model implies that the true regression equation is very similar to the one estimated in the empirical studies. The only difference is that the theory implies that it is σ_γ^2, and not σ_γ, that enters the right hand side of (10). If we use this model to interpret the results of Ramey and Ramey (1995), one must conclude that the negative relationship between mean growth and its standard deviation is evidence that preferences have less curvature than the logarithmic utility, i.e. $0 < \theta < 1$. On the other hand, the Kormendi and Meguire (1985) findings suggest that $\theta > 1$.

In this simple example, the mean investment ratio —the appropriate version of (23)— is

$$z = \frac{\gamma + \delta_k}{\hat{A}}[1 + \sigma_y^2] - \sigma_y^2$$

As was pointed out in the previous section, the covariance between the investment-ratio and the growth rate is positive. In this example, the appropriate version of (25) is

$$\sigma_z = \sigma_y \left(\frac{\rho - (1-\theta)(\hat{A} - \delta_k - \frac{\theta}{2}\sigma_y^2 \hat{A}^2}{\theta} \right) \left(1 + \hat{A}^2\right)^{1/2}.$$

[11] In general, this is not the case.

[12] This is not necessary. In addition to differences in preferences —which we will ignore in this chapter— countries can differ in terms of $(\hat{A}, \delta_k)$ as well.

In this case, the increases in σ_y are associated with increases (decreases) in σ_z if $\theta <$ (>)1. Thus, if $\theta < 1$, the higher the (unobserved) variance of the technology shocks (σ_y^2), the higher the (measured) variances of both the growth rate, σ_γ^2, and the investment rate, σ_z^2, and the lower the mean growth rate. Moreover, in this stochastically singular setting the standard deviation of the growth rate and the investment rate are related (although not linearly). Thus, this simple model is consistent with the findings of Barlevy (2002) that the coefficient of σ_z is estimated to be negative, and that its introduction reduces the significance of σ_γ.

This simple model cannot explain the apparent non-linearity in the relationship between mean and standard deviation of the growth rate process which, according to Fatás (2001), is such that the effect of σ_γ on γ is less negative (and can be positive) for high income countries. In order to account for this fact it is necessary to increase the degree of heterogeneity, and to consider non-linear models.

Finally, the model can be reinterpreted as a multi-country model in which markets are incomplete and the distribution of the domestic shocks —the productivity shocks— is common across all countries.[13] More precisely, consider a market structure in which all countries can trade in a perfectly safe international bond market. In this case —which of course implies that mean growth rates are the same across countries— there is an equilibrium in which all countries choose to hold no international bonds, and the world interest rate is

$$r^* = \hat{A} - \delta_k - \theta\sigma_{y_i}^2\hat{A}^2.$$

If there is a common shock that decreases the variability of every country's technology shocks, this has a positive effect on the "world" interest rate, r^*, and an ambiguous impact on the world growth rate.

3.4.2 Case 2 : A Two Sector (Technology) Model

In the previous model, the variance of the growth rate is exogenous and equal to the variance of the technology shock. This is due, in part, to the assumption that the economy does not have another asset that can be used to diversify risk. In this section we present a very simple two-technology (or two sector) version of the model in which the variance of the growth rate is *endogenously* determined by the portfolio decisions of the representative agent. The main result is that, depending on the source of heterogeneity across countries, the relationship between σ_γ and γ need not be monotone. In particular, and depending on the source of heterogeneity across countries, the model is consistent with increases in σ_γ initially associated with increases in γ, and then, for large values of σ_γ, with decreases in the mean growth rate.

[13] It is possible to allow countries to share the same realization of the stochastic process. Even in this case, the demand for bonds is zero at the conjectured interest rate.

To keep the model simple, we assume that the second technology is not subject to shocks, and we ignore depreciation. Thus, formally, we assume that $\eta_b = \sigma_b = \eta_k = 0$. However, unlike the previous case, the "safe" rate of return r satisfies

$$A - \theta\sigma_k^2 < r < A.$$

This restriction implies that $\alpha^* \ \varepsilon \ (0,1)$, and guarantees that both technologies will be used to produce consumption. Since this model is a special case of the results summarized in (20) (we set the depreciation rates equal to zero for simplicity), it follows that the equilibrium mean growth rate and its variance are given by

$$\gamma = \frac{r-\rho}{\theta} + \left(\frac{A-r}{\theta\sigma_k}\right)^2 \frac{1+\theta}{2}, \tag{27a}$$

$$\sigma_\gamma^2 = \left(\frac{A-r}{\theta\sigma_k}\right)^2. \tag{27b}$$

How can we use the model to interpret the cross country evidence on variability and growth? A necessary first step is to determine which variables can potentially vary across countries. In the context of this example, a natural candidate is the vector (A, r, σ_k). Before we proceed, it is useful to describe the connection between γ and σ_γ implied by the model. The relationship is —taking a discrete time approximation—

$$\begin{aligned} \gamma_t &= \frac{r-\rho}{\theta} + \sigma_\gamma^2 \frac{1+\theta}{2} + \varepsilon_t, \\ \varepsilon_t &= \sigma_\gamma \omega_t, \qquad \omega_t \sim N(0,1). \end{aligned}$$

It follows that if the source of cross-country differences are differences in (A, σ_k) the model implies that —independently of the degree of curvature of preferences— the relationship between σ_γ^2 and γ is positive. To see why increases in σ_k result in such a positive association between the two endogenous variables σ_γ and γ, note that, as σ_k rises, the economy shifts more resources to the safe technology (α^* decreases) and this, in turn, results in a decrease in the variance of the growth rate (which is a weighted average of the variances of the two technologies). Since the 'risky' technology has higher mean return than the 'safe' technology, the mean growth rate decreases. The reader can verify that changes in A have a similar effects.

If the source of cross-country heterogeneity is due to differences in r, the implications of the model are more complex. Consider the impact of a decrease in r. From (27b) it follows that σ_γ^2 increases and this tends to increase γ. However, as (27a) shows, this also decreases the growth rate, as it lowers the non-stochastic return. The total effect depends on the combined impact. A simple calculation shows that

$$\frac{\partial\gamma}{\partial r} \lesseqgtr 0 \qquad \Leftrightarrow \qquad r \lesseqgtr \hat{r},$$

where

$$\hat{r} = A - \frac{\theta \sigma_k^2}{1+\theta}.$$

To better understand the implications of the model consider a "high" value of r; in particular, assume that $r > \hat{r}$. A decrease in r reduces σ_γ and, given that $r > \hat{r}$, it results in an increase in γ. Thus, for low σ_γ (high r) countries, the model implies a positive relationship between γ and σ_γ. If $r < \hat{r}$, decreases in the return to the safe technology still increase σ_γ, but, in this region, the growth rate decreases. Thus, in (σ_γ, γ) space the model implies that, due to variations in r, the relationship between σ_γ and γ has an inverted U-shape.

Can this model explain some of the non-linearities in the data? In the absence of further restrictions on the cross-sectional joint distribution of (A, r, σ_k) the model can accommodate arbitrary patterns of association between σ_γ and γ. If one restricts the source of variation to changes in the return r the model implies that, for high variance countries, variability and growth move in the same direction, while for low variance countries the converse is true. If one could associate low variance countries with relatively rich countries, the implications of the model would be consistent with the type of non-linearity identified by Fatás (2001).

3.4.3 Case 3: Aggregate vs. Sectoral Shocks

The simple Ak model that we discussed in the previous section is driven by a single, aggregate, shock. In this section we consider a two sector (or two technology) economy to show that the degree of sectoral correlation of the exogenous shocks can affect the mean growth rate. To capture the ideas in as simple as possible a model, we specialize the specification in (14) by considering the case

$$\begin{aligned} \sigma_k &= \sigma_b = \sigma > 0, \\ \eta_b &= 0,\ \eta_k = \eta, \\ \delta_k &= \delta_b = 0. \end{aligned}$$

Note that, in this setting, there is an aggregate shock, W_t, which affects both sectors (technologies) while the A sector is also subject to a specific shock, Z_t^k. Using the formulas derived in (18) and (20) it follows that the relevant equilibrium quantities are

$$\begin{aligned} \alpha^* &= \frac{A-r}{\theta \eta^2}, \\ \gamma &= \frac{r-\rho}{\theta} - (1-\theta)\frac{\sigma^2}{2} + \left(\frac{A-r}{\theta\eta}\right)^2 \frac{1+\theta}{2}, \\ \sigma_\gamma^2 &= \sigma^2 + \left(\frac{A-r}{\theta\eta}\right)^2. \end{aligned}$$

As before, it is useful to think of countries as indexed by (A, r, σ, η). Since changes in each of these parameters has a different impact, we analyze them separately.

- `An increase in` σ. The increase in the standard deviation of the economy-wide shock affects both sectors equally, and it does not induce any 'portfolio' or sectoral reallocation of capital. The share of capital allocated to each sector (technology) is independent of σ. Since increases in σ increase σ_γ (in the absence of a portfolio reallocation, this is similar to the one sector case), the total effect of an increase in σ is to decrease the growth rate if $0 < \theta < 1$, and to increase it if $\theta > 1$.

- `A decrease in` r. The effect of a change in r parallels the discussion in the previous section. It is immediate to verify that a decrease in r results in an increase in σ_γ. However, the impact on γ is not monotonic. For high values of r, decreases in r are associated with increases in γ, while for low values the direction is reversed. Putting together these two pieces of information, it follows that the predicted relationship between σ_γ and γ is an inverted U-shape, with a unique value of σ_γ (a unique value of r) that maximizes the growth rate.

- `An increase in` η. This change increases the 'riskiness' of the A technology and results in a portfolio reallocation as the representative agent decreases the share of capital in the high return sector (technology). The change implies that σ_γ and γ decrease. Thus, differences in η induce a positive correlation between mean and standard deviation of the growth rate.

- What is the impact of differences in the degree of correlation between sectoral shocks. Note that the correlation between the two sectoral shocks is

$$\nu = \frac{\sigma}{(\sigma^2 + \eta^2)^{1/2}}.$$

  In order to isolate the impact of a change in correlation, let's consider changes in (σ, η) such that the variance of the growth rate is unchanged. Thus, we restrict (σ, η) to satisfy

$$\sigma_\gamma^2 = \sigma^2 + \left(\frac{A - r}{\theta\eta}\right)^2,$$

  for a given (fixed) σ_γ. It follows that the correlation between the two shocks and the growth rate are

$$\begin{aligned} \nu &= \left(1 + (\frac{A - r}{\theta})^2 \frac{1}{\sigma^2(\sigma_\gamma^2 - \sigma^2)}\right)^{-1}, \\ \gamma &= \frac{r - \rho}{\theta} - \sigma^2 + \frac{1 + \theta}{2}\sigma_\gamma^2. \end{aligned}$$

Thus, lower correlation between sectors (in this case this corresponds to higher σ) unambiguously lower mean growth. If countries differ in this correlation then the implied relationship between σ_γ and γ need not be a function; it can be a correspondence. Put it differently, the model is consistent with different values of γ associated to the same σ_γ.

3.5 Physical and Human Capital

In this section we study models in which individuals invest in human and physical capital. We consider a model in which the rate of utilization of human capital is constant. Even though the model is quite simple it is rich enough to be consistent with **any** estimated relationship between σ_γ and γ.

We assume that output can be used to produce consumption and investment, and that market goods are used to produce human capital. This is equivalent to assuming that the production function for human capital is identical to the production function of general output. The feasibility constraints are

$$\begin{aligned} dk_t &= ([F(k_t, h_t) - \delta_k k_t - x_t - c_t)\, dt + \sigma_y F(k_t, h_t) dW_t, \\ dh_t &= -\delta_h h_t + x_t dt + \sigma_h h_t dW_t + \eta h_t dZ_t, \end{aligned}$$

where (W_t, Z_t) is a vector of independent standard Brownian motion variables, and F is a homogeneous of degree one, concave, function. As in the previous sections, let $x_t = k_t + h_t$ denote total (human and non-human) wealth. With this notation, the two feasibility constraints collapse to

$$\begin{aligned} dx_t &= ([F(\alpha_t, 1-\alpha_t) - (\delta_k \alpha_t + \delta_h(1-\alpha_t))]x_t - c_t)\, dt + \\ \sigma_y F(\alpha_t, 1-\alpha_t) x_t dW_t + & \quad \sigma_h (1-\alpha_t) x_t dW_t + \eta(1-\alpha_t) x_t dZ_t. \end{aligned} \tag{28}$$

As in previous sections, the competitive equilibrium allocation coincides with the solution to the planner's problem. The planner maximizes (13) subject to (28). The Hamilton-Jacobi-Bellman equation corresponding to this problem is

$$\rho V(x) = \max_{c,\alpha} \left[\frac{c^{1-\theta}}{1-\theta} + V'(x)[(F(\alpha, 1-\alpha) - \delta(\alpha))x_t - c_t] + \frac{V''(x)x^2}{2}\sigma^2(\alpha) \right],$$

where

$$\begin{aligned} \delta(\alpha) &= \delta_k \alpha + \delta_h(1-\alpha), \\ \sigma^2(\alpha) &= \sigma_y^2 F(\alpha, 1-\alpha)^2 + \sigma_h^2 (1-\alpha)^2 + \eta^2 (1-\alpha)^2 + \sigma_y \sigma_h F(\alpha, 1-\alpha)(1-\alpha). \end{aligned}$$

It can be verified that a function of the form $V(x) = v\frac{x^{1-\theta}}{1-\theta}$ solves the Hamilton-Jacobi-Bellman equation. The solution also requires that

$$\rho = \theta v^{-1/\theta} + (1-\theta)\{F(\alpha, 1-\alpha) - \delta(\alpha) - \frac{\theta}{2}\sigma^2(\alpha)\},$$

where α is given by

$$\alpha = \arg\max(1-\theta)\{F(\alpha, 1-\alpha) - \delta(\alpha) - \frac{\theta}{2}\sigma^2(\alpha)\}.$$

For any homogeneous of degree one function F, the solution is a constant α. Moreover, α does not depend on v. Existence requires $v > 0$, and this is just a condition on the exogenous parameter that we assume holds.[14]

The growth rate and its variance are given by

$$\begin{aligned} \gamma &= F(\alpha, 1-\alpha) - \delta(\alpha) - v^{-1/\theta}, \\ \sigma_\gamma^2 &= \sigma_y^2 F(\alpha, 1-\alpha)^2 + \sigma_h^2(1-\alpha)^2 + \eta^2(1-\alpha)^2 + \sigma_y\sigma_h F(\alpha, 1-\alpha)(1-\alpha) \end{aligned}$$

It follows that, for the class of economies for which the planner problem has a solution (i.e. economies for which $v > 0$, and $\gamma > 0$), the conjectured form of $V(x)$ solves the HJB equation, for any homogeneous of degree one function F. However, in order to make some progress describing the implications of the theory, it will prove convenient to specialize the technology and assume that F is a Cobb-Douglas function given by

$$F(x, y) = Ax^\omega y^{1-\omega}, \qquad 0 < \omega < 1.$$

The next step is to characterize the optimal share of wealth invested in physical capital, α, and how changes in country-specific parameters affect the mean and standard deviation of the growth rate. It turns out that the qualitative nature of the solution depends on the details of the driving stochastic process. To simplify the algebra, we assume that the human capital technology is deterministic (i.e. $\sigma_h = \eta = 0$), and that both stocks of capital (physical and human) depreciate at the same rate ($\delta_k = \delta_h$). As indicated above, we assume that the production function is Cobb-Douglas. The first order condition for the optimal choice of α is simply

$$\phi(\alpha)\hat{F}(\alpha)[1 - \theta\sigma_y^2\hat{F}(\alpha)] = 0,$$

where

$$\begin{aligned} \hat{F}(\alpha) &\equiv A\alpha^\omega(1-\alpha)^{1-\omega}, \\ \phi(\alpha) &= \frac{\omega}{\alpha} - \frac{1-\omega}{1-\alpha}. \end{aligned}$$

The second order condition requires that

$$-\omega(1-\omega)[\alpha^{-2} + (1-\alpha)^{-2}]\hat{F}(\alpha)[1 - \theta\sigma_y^2\hat{F}(\alpha)] - \theta\sigma_y^2\hat{F}(\alpha)^2\phi(\alpha) < 0.$$

Since $\hat{F}(\alpha) > 0$ in the relevant range, the solution is either $\phi(\alpha) = 0$, which corresponds to $\alpha^* = \omega$, or $\hat{F}(\alpha^*) = 1/\theta\sigma_y^2$. The latter, of course, does not result in a unique α^*[15].

[14] This is just the stochastic analog of the existence problem in endogenous growth models.

[15] In the case of the Cobb-Douglas production function there are two values of α that satisfy $\hat{F}(\alpha^*) = 1/\theta\sigma_y^2$

The nature of the solution depends on the size of σ_y^2. There are two cases characterized by

- *Case A*: $\sigma_y^2 \leq \frac{1}{\theta \hat{F}(\omega)}$.
- *Case B*: $\sigma_y^2 > \frac{1}{\theta \hat{F}(\omega)}$.

In *Case A*, the low variance σ_y^2 case, the maximizer is given by $\alpha^* = \omega$, since $1 - \theta\sigma_y^2 \hat{F}(\alpha) > 0$ for all feasible α. The second order condition is satisfied.

In *Case B*, there are two solutions to the first order condition. They correspond to the values of α, denoted α^- and α^+ that solve $\hat{F}(\alpha^*) = 1/\theta\sigma_y^2$. By convention, let's consider $\alpha^- < \omega < \alpha^+$. It can be verified that in both cases the second order condition is satisfied.[16] The implications of the model for the expected growth rate and its standard deviation in the two cases are

$$\begin{aligned}
\gamma_A &= \frac{\hat{F}(\omega) - (\rho + \delta)}{\theta} - \frac{1-\theta}{2}\sigma_y^2 \hat{F}(\omega)^2, \\
\sigma_{\gamma_A} &= \sigma_y \hat{F}(\omega), \\
\gamma_B &= \frac{1}{\theta}[\frac{1+\theta}{2}\frac{1}{\theta\sigma_y^2} - (\rho + \delta)], \\
\sigma_{\gamma_B} &= \frac{1}{\theta\sigma_y}.
\end{aligned}$$

It follows that for large σ_y^2, that is in *Case B*, the model predicts a positive relationship between mean growth and the standard deviation of the growth rate, while for small values of σ_y^2, *Case A*, the sign of the relationship depends on the magnitude of θ.

Much more interesting from a theoretical point of view is the fact that the model is consistent with two countries with different σ_y^2 to have *exactly the same* σ_γ. To see this, note that for any σ_γ in the range of feasible values —corresponding to the set $[0, \left(\frac{\hat{F}(\omega)}{\theta}\right)^{1/2}]$ in this example— there are two values of σ_y, one less than $\left(\frac{1}{\theta\hat{F}(\omega)}\right)^{1/2}$, and the other greater than this threshold that result in the same σ_γ. The relationship between σ_γ and γ is a correspondence. Figure 1 displays such a relationship in the small risk aversion case, $0 < \theta < 1$.

If the only source of cross-country heterogeneity are differences in the variability of the technology shocks, σ_y, the model implies that all data points should be in one of the two branches of the mapping depicted in Figure 1. By arbitrarily choosing the location of these points, the estimated relationship between σ_γ and γ can have any sign, and the estimated value says very little about the deep parameters of the model

[16] The reader can check that, in this case, the solution $\alpha^* = \omega$ does not satisfy the second order condition.

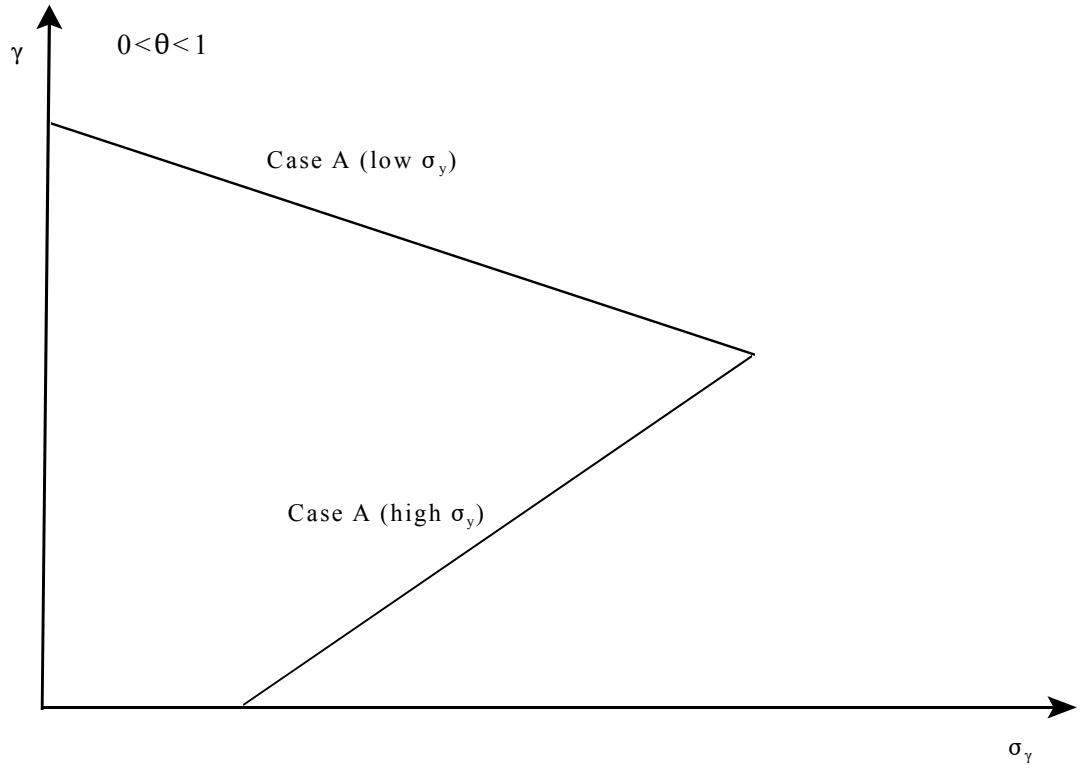


Figure 1: Figure 1: The mapping between σ_γ and γ. $[0 < \theta < 1]$

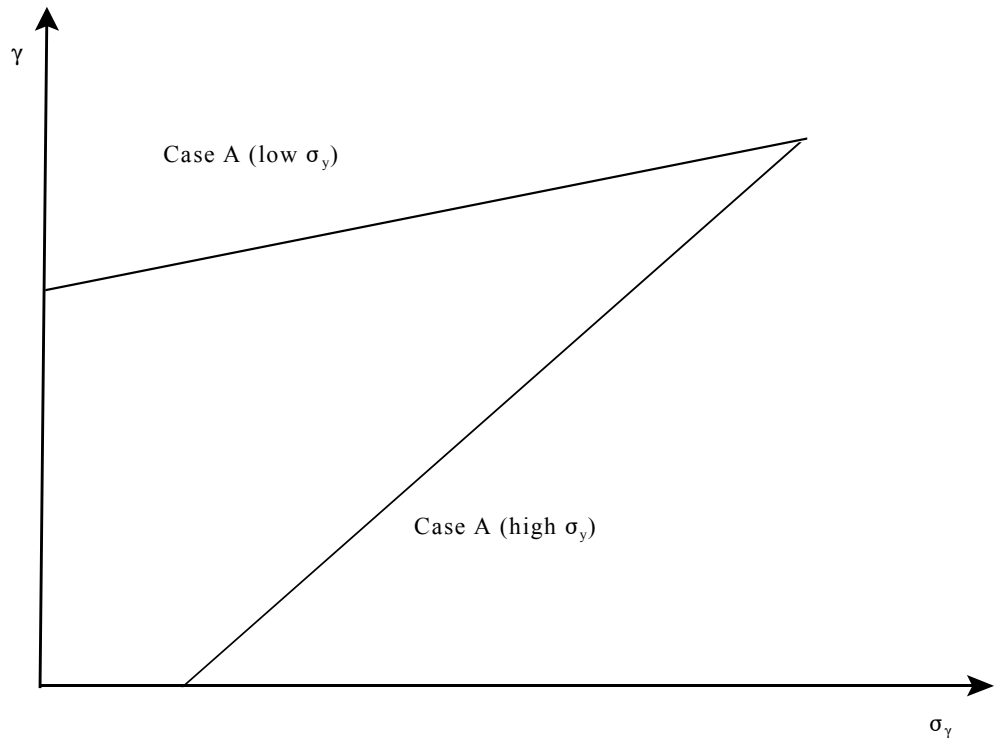


Figure 2: Figure 2: The mapping between σ_γ and γ. $[\theta > 1]$

or, more importantly, about the effects of reducing the variability of shocks on the average growth rate.

Does the nature of the result depend on the assumption $0 < \theta < 1$? For $\theta > 1$ the relationship between σ_γ and γ is also a correspondence, and hence that the model —in the absence of additional assumptions— does not pin down the sign of the correlation between σ_γ and γ.[17] In the case of $\theta > 1$, the size of θ matters only to determine which branch is steeper. I both *cases* A and B the relationship between the standard deviation of the growth rate and the mean growth rate is upward sloping. However, the low σ_y^2-branch is flatter (and lies above) the high σ_y^2-branch (see Figure 2).

Since we have studied a very simple version of this class of models, it does not seem useful to determine the relevance of each branch by asigning values to the parameters. In ongoing work, we are studying more general versions of this setup. However, even this simple example suggests that some caution must be exercised when interpreting the empirical work relating the variability of the growth rate and its mean. Unless

[17] At this point, we have not explored what are the consequences of adding the investment output ratio to the (theoretical) regression. However, to do this is a complete manner it seems necessary to model measurement errors, as the model is stochastically singular. We leave this for future work.

one can rule out some of these cases, theory gives ambiguous answers to the question that motivated much of the literature, i.e. Do more stable countries grow faster? Moreover, the theoretical developments suggest that progress will require to estimate structural models rather than reduced form equations.

3.6 The Opportunity Cost View

So far the models we discussed emphasize the idea that increases in the variability of the driving shocks can have positive or negative effects upon the growth rate depending on the relative importance of income and substitution effects. An alternative view is that recessions are "good times' to invest in human capital because labor —viewed as the single most important input in the production of human capital— has a low opportunity cost. In this section we present a model that captures these ideas. The model implies that the time allocated to the formation of human capital is independent of the cycle.[18] It also implies that shocks to the goods production technology have no impact on growth, but that the variability of the shock process in the human capital technology decreases growth.

As before, we concentrate on a representative agent with preferences described by (13). The goods production technology is given by

$$c_t + x_t \leq z_t A k_t^{\alpha} (n_t h_t)^{1-\alpha},$$

where n_t is the fraction of the time allocated to goods production, k_t is the stock of physical capital, and h_t is the stock of human capital. The variable z_t denotes a stationary process. To simplify the theoretical presentation we assume that capital depreciates fully. Thus, goods consumption is limited by

$$c_t \leq z_t A k_t^{\alpha} (n_t h_t)^{1-\alpha} - k_t.^{19}$$

Human capital is produced using only labor in order to capture the idea that the opportunity cost of investing in human capital is market production. The technology is summarized by

$$dh_t = [1 - \delta + B(1 - n_t)] h_t dt + \sigma_h [1 - \delta + B(1 - n_t)] h_t dW_t,$$

[18] The empirical relationship between investment in human capital and the cycle is mixed. Dellas and Sakellaris (1997) using CPS data for all individuals aged 18 to 22 find that college enrollment id procyclical. Christian (2002) also using the CPS but restricting the sample to 18-19 years olds (sa as to be able to control for family variables) finds no cyclical effects. Sakellaris and Spilimbergo (2000) study U.S. college enrollment of foreign nationals and conclude that, among those individuals coming from risch countries enrollment is countercyclical, while among students from less developed countries it is countercyclical. Moreover, college enrollment is only a partial measure of investment in human capital. Training (inside and outside business firms) is another (difficult to measure) component of increases in skill acquisition.

[19] This restriction makes it possible to derive the theoretical implications of the model in a simple setting.

where, as before, W_t is a standard Brownian motion.[20]

Given that the problem is convex[21] the competitive allocation solves the planner's problem. It is clear that, given $n_t h_t$, physical capital will be chosen to maximize net output. This implies that consumption is

$$c_t = A^* \hat{z}_t n_t h_t,$$

where $A^* = (A\alpha)^{1/(1-\alpha)}(\alpha^{-1} - 1)$ and $\hat{z}_t = z_t^{1/(1-\alpha)}$. We guess that the relevant state variable is the vector $(\hat{z}_t, h_t)$, and that the value function is of the form

$$V(\hat{z}_t, h_t) = v\frac{(\hat{z}_t h_t)^{1-\theta}}{1-\theta}.$$

Given this guess, the relevant Hamilton-Jacobi-Bellman equation is

$$\rho v \frac{(\hat{z}h)^{1-\theta}}{1-\theta} = \max_x \left\{ \frac{[\frac{A^*}{B}(\mu - x)\hat{z}h]^{1-\theta}}{1-\theta} + v(\hat{z}h)^{1-\theta}x - v(\hat{z}h)^{1-\theta}\theta\frac{\sigma_h^2}{2}x^2, \right\}$$

where $\mu \equiv 1-\delta+B$, and $x = 1-\delta+B(1-n)$. It follows that choosing x is equivalent to choosing n. The solution to the optimization problem is given by the solution to the following quadratic equation

$$x^2 = \frac{2(1+\mu\sigma_h^2)}{(1+\theta)\sigma_h^2}x + \frac{2(\rho-\mu)}{\theta(1+\theta)\sigma_h^2}.$$

In order to guarantee that utility remains bounded even in the case $\sigma_h = 0$ is is necessary to assume that $\rho - \mu > 0$. Simple algebra shows that the positive root of the previous equation is such that increases in σ_h decrease x. It follows that the stochastic process for h_t is given by

$$dh_t = xh_t dt + \sigma_h h_t dW_t$$

We now discuss the implications of the model for the growth rate of consumption (or output). Even though our results do not depend on the particular form of the z_t process, it is convenient to consider the case in which z_t is a geometric Brownian motion that is possibly correlated with the shock to the human capital. Specifically, we assume that

$$dz_t = z_t(\sigma_w dW_t + \sigma_m dM_t),$$

where M_t is a standard Brownian motion that is uncorrelated with W_t. Ito's lemma implies that

$$d\hat{z}_t = \frac{\alpha}{(1-\alpha)^2}\frac{\sigma_w^2+\sigma_m^2}{2}\hat{z}_t dt + \frac{\alpha}{(1-\alpha)}\hat{z}_t(\sigma_w dW_t + \sigma_m dM_t).$$

[20] A special case of this model in which utility is assumed logarithmic, and the goods production function is not subject to shocks is analyzed in De Kek (1999).

[21] Even though our choice of notation somewhat obscures this, the convexity of the technology is apparent by defining $h_{mt} = n_t h_t$ and $h_{st} = (1-n_t)h_t$, and adding the constraint $h_{mt} + h_{st} \leq h_t$.

In equilibrium, consumption (and net output) is given by

$$c_t = \frac{A^*}{B}(\mu - x)\hat{z}_t h_t.$$

Applying Ito's lemma to this expression, we obtain that the growth rate of consumption

$$\frac{dc_t}{c_t} = \frac{dh_t}{h_t} + \frac{d\hat{z}_t}{\hat{z}_t} + \frac{\alpha x}{(1-\alpha)}\sigma_h \sigma_w dt,$$

or, taking a discrete time approximation,

$$\begin{aligned}
\gamma_t &= x(1 + \frac{\alpha}{(1-\alpha)}\sigma_h\sigma_w) + \frac{\alpha}{(1-\alpha)^2}\frac{\sigma_w^2 + \sigma_m^2}{2} + \\
&\quad [(\frac{\alpha}{(1-\alpha)}\sigma_w + \sigma_h x)\tilde{W}_t + \frac{\alpha}{(1-\alpha)}\sigma_m \tilde{M}_t], \qquad (29a)\\
\gamma_t &= \gamma + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma_\gamma^2), \qquad (29b)\\
\sigma_\gamma^2 &= (\frac{\alpha}{(1-\alpha)}\sigma_w + \sigma_h x)^2 + (\frac{\alpha}{(1-\alpha)}\sigma_m)^2 \qquad (29c)
\end{aligned}$$

Equation (29a) completely summarizes the implications of the model for the data. There are several interesting results. To simplify the notation, we will refer to W_t as the aggregate shocks and to M_t as the idiosyncratic component of the productivity shock in the goods sector.

- The share of the time allocated to human capital formation —the engine of growth in this economy— is independent of the variability of the technology shock in the goods sector, as measured by (σ_w, σ_m).

- High (σ_w, σ_m) economies are also high growth economies. Thus, if cross-country differences in σ_γ are mostly due to differences in (σ_w, σ_m), the model implies a positive correlation between the standard deviation of the growth rate and mean growth.

- It can be shown that increases in σ_h result in *decreases* in $\sigma_h x$. Thus, if countries differ in this dimension the model also implies a positive relationship between σ_γ and γ.

- In the model, investment in physical capital (as a fraction of output) is α, independently of the distribution of the shocks. Thus, there is no sense that a regression that shows that variability does not affect the rate of investment provides evidence against the role of shocks in development.

- This lack of (measured) effect on both physical and human capital investment should not be interpreted as evidence against the proposition that incentives for human or physical capital accumulation matter for growth. It is easy enough to include a tax/subsidy to the production of human capital —consider a policy that affects B— and it follows that this policy affects growth.

3.7 More on Government Spending, Taxation, and Growth

In this section we consider a simple Ak model in which a government uses distortionary taxes to finance an exogenously given stochastic process for government spending. Our analysis follows Eaton (1981).[22]

The representative household maximizes utility —given by (13)— by choosing consumption and saving in either capital or bonds. However, given that tax policy is exogenously fixed, it is not the case that the rate of return on bonds is risk free. On the contrary, since the government issues bonds to make up for any difference between revenue and spending it is necessary to let the return on bonds to be stochastic.

The representative household problem is

$$\max U = E\left[\int_0^\infty e^{-\rho t}\frac{c_t^{1-\theta}}{1-\theta}dt\,|F_0\right]. \tag{30}$$

subject to

$$\begin{aligned} dk_t &= (r_k k_t - c_{1t})dt + \sigma_k k_t dW_t, && (31a)\\ db_t &= (r_b b_t - c_{2t})dt + \sigma_b b_t dW_t, && (31b)\\ c_t &= c_{1t} + c_{2t}, && (31c) \end{aligned}$$

where k_t is interpreted as capital and b_t as bonds. As before, it is possible to simplify the analysis by using wealth as the state variable. Let $x_t \equiv k_t + b_t$. With this notation, the single budget constraint is given by,

$$dx_t = [(\alpha_t r_k + (1-\alpha_t)r_b)x_t - c_t]dt + (\alpha_t\sigma_k + (1-\alpha_t)\sigma_b)x_t dW_t.$$

Since this problem is a special case of the "general" two risky assets model, it follows that the optimal solution is characterized by

$$\begin{aligned} \alpha &= \frac{\frac{r_k - r_b}{\theta} - \sigma_b(\sigma_k - \sigma_b)}{(\sigma_k - \sigma_b)^2}, && (32a)\\ c_t &= \underbrace{\frac{\rho - (1-\theta)[\alpha r_k + (1-\alpha)r_b - \theta\frac{(\alpha\sigma_k + (1-\alpha)\sigma_b)^2}{2}]}{\theta}}_{c} x_t. && (32b) \end{aligned}$$

The set of feasible allocations is the set of stochastic process that satisfy

$$\begin{aligned} dk_t &= (Ak_t - c_t)dt + \sigma A k_t dW_t - dG_t,\\ dG_t &= gAk_t dt + g'\sigma A k_t dW_t. \end{aligned}$$

Thus, the government consumes a fraction g of the non-stochastic component of output, and a fraction g' of the stochastic component. Taxes are levied on the

[22] For extensions of this model, see Turnovski (1995)

deterministic and stochastic components of output at (possibly) different rates. The stochastic process for tax revenue is assumed to satisfy

$$dT_t = \tau A k_t dt + \tau' \sigma A k_t dW_t.$$

In equilibrium, the parameters that determine the rate of return on capital (r_k, σ_k) are given by

$$\begin{aligned} r_k &= (1-\tau)A, \\ \sigma_k &= (1-\tau')\sigma A. \end{aligned}$$

The government budget constraint requires that the excess of spending over tax revenue be financed through bond issues. Thus,

$$B_t + dG_t - dT_t = p_t dB_t,$$

where $p_t B_t = b_t$ is the value of bonds issued. The stock of capital evolves according to

$$dk_t = ((1-g)A - \frac{c_t}{k_t})k_t dt + \sigma(1-g')Ak_t dW_t.$$

Note that

$$\frac{c_t}{k_t} = c\frac{x_t}{k_t} = c(1 + \frac{1-\alpha}{\alpha}) = \frac{c}{\alpha}.$$

Since, in equilibrium, it must be the case that, in all states of nature, the growth rate of private wealth and the growth rate of the capital stock are the same[23], it is necessary that

$$\alpha r_k + (1-\alpha)r_b - c = (1-g)A - \frac{c}{\alpha}, \tag{33a}$$

$$\alpha\sigma_k + (1-\alpha)\sigma_b = \sigma(1-g')A. \tag{33b}$$

The system formed by the four equations described in (32) and (33) provides the solution to the endogenous variables that need to be determined: c, α, r_b, σ_b. It is convenient to define the excess rate of return of capital, and the excess instant variability of capital as

$$\begin{aligned} \Delta_r &= r_k - r_b, \\ \Delta_\sigma &= \sigma_k - \sigma_b. \end{aligned}$$

Some simple but tedious algebra shows that

$$\begin{aligned} \alpha &= \frac{\sigma_\gamma - \sigma_k + \Delta_\sigma}{\Delta_\sigma}, \\ \Delta_r &= \theta\sigma_\gamma\Delta_\sigma \end{aligned}$$

[23] This, of course, depends on the fact that the solution to the individual agent problem is such that bonds and capital are held in fixed proportions.

Substituting in the remaining equations, and recalling that the instantaneous mean and standard deviation of the growth rate process is given by

$$\begin{aligned} \gamma &= \alpha r_k + (1-\alpha) r_b - c, \\ \sigma_\gamma &= \sigma(1-g')A, \end{aligned}$$

it follows that,

$$\gamma = \frac{(1-\tau)A - \rho}{\theta} - \left(\frac{1-\theta-\tau'+\theta g'}{1-g'} \right) \sigma_\gamma^2 \tag{34a}$$

$$\sigma_\gamma = \sigma(1-g')A. \tag{34b}$$

Equation (34a) summarizes the impact of both technology and fiscal shocks on the expected growth rate. Consider first the impact of variations in the tax regime on the relationship between the variability of the growth rate, σ_γ, and the average growth rate. If technology shocks, σ, are the main source of differences across countries in the standard deviation of the growth rate, then high variability countries are predicted to be low mean countries if $1-\theta-\tau'+\theta g' > 0$; that is, if a country has a relatively low tax rate on the stochastic component of income. This, would be the case if the base of the income tax allowed averaging over several periods. On the other hand, countries with relatively high tax rates on the random component of income display a positive relationship between the mean and the variance of their growth rates.

As in more standard models, high capital income tax countries (high τ countries) have lower average growth. Differences across countries in the average size of the government, g in this notation, have no impact on growth. Finally, cross country differences in the fraction of the random component of income consumed by the government, g', induce a positive correlation between γ and σ_γ. This is driven by the negative impact that increases in g' have on mean growth, and the equally negative effect that those changes have on σ_γ. Thus, high g' countries display low average growth rates, which do not fluctuate much.[24]

3.8 Quantitative Effects

In this section we summarize some of the quantitative implications various models for the relationship between variability and growth. Unlike the theoretical models described above, the quantitative exercises concentrate on the role of technology shocks in models with constant –relative to output– government spending.

[24] The impact of some variables in the previous analysis differs from the results in Eaton (1981) since our specification of the fiscal policy allows the demand for bonds (as a fraction of wealth) to be endogenous, and driven by changes in the tax code. Eaton assumes that the share of bonds, $1-\alpha$ in our notation, is given, and some tax must adjust to guarantee that demand and supply of bonds are equal.

Mendoza (1997) studies an economy in which the planner solves the following problem

$$\max_{\{c_t\}} \sum_{t=0}^{\infty} \beta^t \frac{c_t^{1-\theta}}{1-\theta}$$

subject to

$$A_{t+1} \leq R_t(A_t - p_t c_t),$$

where A_t is a measure of assets and p_t is interpreted as the terms of trade. Since this equation can be rewritten as

$$k_{t+1} = r_{t+1}(k_t - c_t),$$

where $k_t = A_t/p_t$ is the stock of assets (capital) measured in units of consumption, and $r_{t+1} = R_t p_t/p_{t+1}$ is the random rate of return, it is clear that Mendoza's model is a stochastic version of an Ak model. The rate of return, r_t is assumed to be lognormally distributed with mean and variance given by

$$\begin{aligned} \mu_r &= e^{\mu+\sigma^2/2}, \\ \sigma_r^2 &= \mu_r^2 e^{\sigma^2-1}. \end{aligned}$$

It follows that the standard deviation of the growth rate is $\sigma_\gamma = \sigma$. Mendoza studies the effect of changing σ from 0 to 0.15, holding μ_r constant. To put the exercise in perspective, the average across countries of the standard deviations of the growth rate or per capita output in the Summers-Heston dataset is 0.06. Thus, the model is calibrated at a fairly high level of variability. The results depend substantially on the assumed value of θ. For $\theta = 1/2$, the non-stochastic growth rate is 3.3%. If $\sigma = 0.10$, it decreases to 2.5%, while it is 1.6% when $\sigma = 0.15$. For $\theta = 2.33$ (Mendoza's preferred specification), the growth rate increases from 0.7% to 0.9% in the given range. For other values of the coefficient of risk aversion, the impact of uncertainty is also small. In summary, unless preferences are such that the degree of intertemporal substitution is large, increases in rate of return uncertainty have a small impact on mean growth rates.

Jones et. al (2003a) analyze the following planner's problem:

$$\max E_t\{\sum_t \beta^t c_t^{1-\theta} v(\ell_t)/(1-\theta)\},$$

subject to,

$$\begin{aligned} c_t + x_{zt} + x_{ht} + x_{kt} &\leq F(k_t, z_t, s_t), \\ z_t &\leq M(n_{zt}, h_t, x_{zt}) \\ k_{t+1} &\leq (1-\delta_k)k_t + x_{kt}, \\ h_{t+1} &\leq (1-\delta_h)h_t + G(n_{ht}, h_t, x_{ht}) \\ \ell_t + n_{ht} + n_{zt} &\leq 1. \end{aligned}$$

For their quantitative exercise, they specify the following functional forms

$$\begin{aligned} n_h &= x_z = 0, \quad n_z = n, \\ v(\ell) &= \ell^{\psi(1-\sigma)}, \\ F(k,z,s) &= sAk^{\alpha}z^{1-\alpha}, \\ G(h,x_h) &= x_h, \\ M(n,h) &= nh, \\ s_t &= \exp[\zeta_t - \frac{\sigma_\varepsilon^2}{2(1-\rho^2)}], \\ \zeta_{t+1} &= \rho\zeta_t + \varepsilon_{t+1}. \end{aligned}$$

The model is calibrated to match the average growth, in a cross section of countries, and its standard deviation. Jones et. al (2003) consider the impact of changing the standard deviation of the shock, s_t, from 0 to 0.15. The impact on the growth rate depends on the curvature of the utility function. For preferences slightly less curved than the log, the model predicts an increase in the growth rate of 0.7% an an annual basis, while for $\theta = 1.5$, the effects is an increase in the growth rate of 0.25%. However, the model predicts that σ_γ —which is endogenous— is unusually high (of the order of 0.10) unless $\theta \geq 1.5$.

Thus, Jones et. al (2003a) obtain results that are quite different from those of Mendoza (1997). There are two important differences between the models: First, while Mendoza (1997) assumes that shocks are *i.i.d.*, Jones et. al. (2003a) set the first order correlation parameter, ρ, to 0.95. Second, while Mendoza assumes a constant labor supply, Jones et. al allow for the number of hours to vary with the shock.

In order to disentangle the effect of the components of the standard deviation of the technology shock, Jones et. al. vary σ_ε and ρ in a series of experiments, where $\sigma_s = \sigma_\varepsilon/(1-\rho^2)^{1/2}$, where σ_ε is the standard deviation of the innovations. They find that changes in σ_ε appear quantitatively more important than those and ρ. Moreover, they also find that the relative variability of hours worked is very sensitive to the precise value of θ. Economies with high θ, and lower σ_γ in their specification, also display substantially less variability in hours worked. Even though it is not possible to determine on the basis of these two results which is the critical feature that accounts for the differences between the results obtained by Mendoza (1997) and Jones et. al (2003a), it seems that the assumption of a flexible supply of hours —which determines the rate of utilization of human capital— is a leading candidate.[25]

In a series of papers, Krebs (2003) and (2004) explores the impact of changes in uncertainty in models where markets are incomplete. Building on the work of de Hek (1999), Krebs (2003) studies the impact of shocks to the depreciation rate of the

[25] In subsequent work, Jones et. al (2003a) analyze the business cycle properties of the same class of models, and they show that are capable of generating higher serial correlation in the growth rate of output than similar exogneous growth models.

capital stock. Even though he assumes that instantaneous utility is logarithmic, he finds that increases in the standard deviation of the shock, decrease growth rates. This result is driven by the "location" of the shocks, and it does not require market incompleteness.[26] Quantitatively, Krebs (2003) finds that an increase in the standard deviation of the growth rate from 0 to 0.15 (a fairly large value relative to the world average) reduces growth from 2.13% to 2.00%. If the variability is increased to 0.20. the growth rate drops to 1.5%. However, these values are substantially higher than those observed in international data.[27]

The theoretical analysis of the impact of different forms of uncertainty on the growth of the economy is still in its infancy. Simple existing suggest that the sign of the relationship between the variability of a country's growth rate and its average growth rate is ambiguous. Thus, theoretical models that restrict more moments can help in understanding the effect of fluctuations on growth and welfare. The few quantitative studies that we reviewed have produced conflicting results. It seems that the precise nature of the shocks, their serial correlation properties and the elasticity of hours with respect to shocks all play a prominent role in accounting for the variance in predicted outcomes. Much more work is needed to identify realistic and tractable models that will be capable of confronting both time series and cross country observations.

4 Concluding Comments

In this chapter we briefly presented the basic insights about the growth process that can be learned from studying standard convex models with perfectly functioning markets. We emphasized three aspects of those models. First, the impact of fiscal policy on growth. A summary of the current state of knowledge is that theoretical models have ambiguous implications about the effect of taxes on growth. The key feature is the importance of market goods in the production of human capital. If, as Lucas (1988) assumes, no market goods are needed to produce new human capital, the impact of income taxes on growth is small (or zero in some cases). If, on the other hand, market goods are necessary to produce human capital then taxes play a more important role, and they have a large impact on growth. It seems that the next step is to use detailed models of the process of human capital formation and to explore the implications that they have for the age-earnings profile to identify the parameters of the production function of human capital. A first step in that direction is in Manuelli and Seshadri (2004).

A second important issue that features prominently in the discussion of the relative

[26] de Hek (1999) shows theoretically that increases in the variance of the depreciation shock decrease average growth even if markets are complete, and the shocks are aggregate shocks.

[27] Krebs (2003) does not have an aggregate shock. His model predicts that aggregate growth is constant. He calibrates his model to match the standard deviation of the growth rate of individual income.

merits of convex and non-convex models is the role of innovation. The standard argument claims that innovation is a one-off investment (with low copying costs) and hence that this technology is inconsistent with price taking behavior. In this chapter we elaborate on the ideas discussed by Boldrin and Levine (2002) and show that it is possible to reconcile the existence of a non-convexity with competitive behavior.

The last major theme covered in this survey is the relationship between fluctuations and growth. An important question is whether technological or policy induced fluctuations affect the growth rate of an economy. This is relevant for the time series experience of a single country(e.g. the discussion about the role of post-war stabilization policies on the growth rate of the American economy), as well as the prescriptions of international agencies for national policies. We discuss the empirical evidence and find it conflicting. It is not easy to identify a clear pattern between fluctuations and growth. To shed light on why this might be the case, we discuss a series of theoretical models. We show that the relationship between the growth rate and its standard deviation has an ambiguous sign. We also describe more precisely how one might identify the parameters of preferences and technologies that determine the sign of the relationship. This is one area of research in which more theoretical and empirical work will have a high marginal value.

References

[1] Aghion, P. and P. Howitt, 1998, **Endogenous Growth Theory**, MIT Press, Cambridge, Massachusetts.

[2] Aizenman, J. and N. Marion, 1999, "Volatility and Investment," *Economica*, 66 (262). pp:

[3] Alvarez, F. and N.L. Stokey, 1998, "Dynamic Programming with Homogeneous Functions," Journal of Economic Theory, Vol. 82, No. 1, pp: 167-189.

[4] Barlevy, G., 2002, "The Cost of Business Cycle Under Endogenous Growth," Northwestern University, working paper.

[5] Barro, R., 1990, "Government Spending in a Simple Model of Economic Growth," *Journal of Political Economy*, Vol 98, Number 5, Part 2, S103-S125.

[6] Barro, R. M and C. Sahasakul, 1986, "Measuring Average Marginal Tax Rates from Social Security and the Individual Income Tax," *Journal of Business*, 59 (4), pp: 555-66.

[7] Barro, R. and X. Sala-i-Martin, 1995, **Economic Growth**, McGraw-Hill, New York, Saint Louis.

[8] Bean, C.I., 1990, "Endogenous Growth and the Procyclical Behavior of Productivity," European Economic Review, 34, pp:355-363.

[9] Bond, E., Ping W. and C. K. Yip, 1996, "A General Two Sector Model of Endogenous Growth with Human and Physical Capital: Balanced Growth and Transitional Dynamics," Journal of Economic Theory, Vol. 68, No. 1, pp: 149-173.

[10] Boldrin, M. and D. Levine, 2002, "Perfectly Competitive Innovation," Federal Reserve Bank of Minneapolis, Staff Report 303.

[11] Brock W. and L. Mirman, 1972, "Optimal Economic Growth Under Uncertainty: The Discounted case," Journal of Economic Theory, 4, pp: 497-513.

[12] Cass, D., 1965, Optimum Growth in an Aggregative Model of Capital Accumulation, Review of Economic Studies, 32, 233-240

[13] Christian, M. S., 2002, "Liquidity Constraints and the Cyclicality of College Enrollment," University of Michigan, working paper.

[14] Cooley, T.F. (ed.), 1995, Frontiers in Business Cycle Research, Princeton University Press, Princeton, New Jersey.

[15] Cooley, T.F and, E. C. Prescott, 1995, "Economic Growth and Business Cycles," in Cooley, T.F. (ed.), Frontiers in Business Cycle Research, Princeton University Press, Princeton, New Jersey, pp: 1-38.

[16] Dawson, J. W. and E. F. Stephenson, 1997, "The Link Between Volatility and Growth: Evidence from the States," Economics Letters, 55, pp: 365-69.

[17] Debreu, G., 1954, "Valuation Equilibria and Pareto Optimum," Proceedings of the National Academy of Sciences, Vol. 40, No. 7, pp: 588-592.

[18] de Hek, P. A., 1999, "On Endogenous Growth Under Uncertainty," International Economic Review, Vol. 40, No.3, pp: 727-744.

[19] de Hek, P. and S. Roy, 2001, "Sustained Growth Under Uncertainty," International Economic Review, Vol. 42, No. 3, pp: 801-813.

[20] Dellas, H. and P. Sakellaris, 1997, "On the Cyclicality of Schooling: Theory and Evidence," University of Maryland, working paper.

[21] Dotsey, M and P-D Sarte, 1997, "Inflation Uncertainty and Growth in a Simple Monetary Model," Federal Reserve Bank of Richmond, May

[22] Eaton, J., 1981, "Fiscal Policy, Inflation and the Accumulation of Risky Capital," Review of Economic Studies, XLVIII, 435-445.

[23] Fatás, A, 2001, "The Effect of Business Cycle on Growth," working paper, INSEAD.

[24] Grier, K. B. and G. Tullock, 1989, "An Empirical Analysis of Cross-National Economic Growth, 1951-1980," Journal of Monetary Economics, 24 (2), pp: 259-76.

[25] Hendricks, L., 2001, "Growth, Taxes and Debt," *Review of Economic Dynamics*, 4(1), pp: 26-57.

[26] Jones, L. E. and R. E. Manuelli, 1990, "A Convex Model of Equilibrium Growth: Theory and Policy Implications," Journal of Political Economy, 98, 1008-1038.

[27] Jones, L. E., R. E. Manuelli and P. E. Rossi, 1993, "Optimal Taxation in Models of Endogenous Growth," Journal of Political Economy, Vol. 101, No. 3, 485-517.

[28] Jones, L. E. and R. E. Manuelli, 1997, "The Sources of Growth" J*ournal of Economic Dynamics and Contro*l, 27, pp: 75-114.

[29] Jones, L. E. and R. E. Manuelli, 1999, "The Equivalence Between Productivity and Tax Shocks," working paper.

[30] Jones, L. E., R. E. Manuelli, H. Siu and E. Stacchetti, 2003a, "Fluctuations in Convex Models of Endogenous Growth I: Growth Effects " working paper.

[31] Jones, L. E., R. E. Manuelli, and H. Siu, 2003b, "Fluctuations in Convex Models of Endogenous Growth II: Business Cycle Properties," working paper.

[32] Judd, K., 1987, "Useful Planning Equivalents of Taxed Economies," working paper.

[33] Judson, R. and A. Orphanides, 1996, "Inflation, Volatility and Growth," Finance and Economics Discussion series 96-19, Federal Reserve Board, May.

[34] Kaganovich, Michael, 1998, "Sustained Endogenous Growth with Decreasing Returns and Heterogeneous Capital," Journal of Economic Dynamics and Control, Vol. 22, Issue 10, pp: 1575-1603.

[35] King, R. G. and S. Rebelo, 1988, "Business Cycles with Endogenous Growth," Rochester Working Paper.

[36] Koopmans, T., 1965, On the Concept of Optimal Economic Growth, in **The Economic Approach to Development Planning**, (North Holland, Amsterdam).

[37] Kormendi, R. L. and P.G. Meguire, 1985, "Macroecomic Determinants of Growth: Cross-Country Evidence," Journal of Monetary Economics, Vol. 16, September, pp: 141-163.

[38] Krebs, T, 2003, "Human Capital Risk and Economic Growth," Quarterly Journal of Economics, 118, pp: 709-745.

[39] Krebs, T, 2004, "Growth and Welfare Effects of Business Cycles in Economies with Idiosyncratic Human Capital Risk," Review of Economic Dynamics (in press)

[40] Kroft, K and Lloyd-Ellis, H, 2002, "Further Cross-Country Evidence on the Link Between Growth, Volatility and Business Cycles," working paper.

[41] Ladron de Guevara, , S. Ortigueira and M. Santos, 1997, "Equilibrium Dynamics in Two-Sector Models of Endogenous Growth," Journal of Economic Dynamics and Control, Vol. 21, No. 1, pp: 115-143.

[42] Leland, H., 1974, "Optimal Growth in a Stochastic Environment," Review of Economic Studies, Vol. 41, N0. 1, pp: 75-86

[43] Levhari D. and T.N. Srinivasan, 1969, "Optimal Savings Under Uncertainty," Review of Economic Studies, Vol XXXVI, No. 106, April, pp: 153-163.

[44] Li, Wenli, and P. Sarte, 2001, "Growth Effects of Progressive Taxes," Federal Reserve Bank of Richmond working paper.

[45] Lucas, R. E., Jr., 1988, "On the Mechanics of Economic Development," Journal of Monetary Economics, 22, 3-42.

[46] Lucas, R. E., Jr., 1990, "Supply-Side Economics: An Analytical Review," *Oxford Economics Papers*, 42, pp: 293-316.

[47] Manuelli, R. and T. J. Sargent, 1988, "Models of Business Cycles: A Review Essay," in Journal of Monetary Economics, 22, pp: 523-542.

[48] Manuelli, R. and A. Seshadri, 2004, "The Contribution of Human Capital to Development," University of Wisconsin, working paper.

[49] Martin, P. and C. A. Rogers, 2000, "Long Term Growth and Short Term Economic Instability," European Economic Review, (44), vol.2, pp: 359-381.

[50] McGrattan, E. and E. C. Prescott, 2003, "Average Debt and Equity Returns: Puzzling?" Federal Reserve Bank of Minneapolis, Staff Report 313.

[51] McGrattan, E. and E. C. Prescott, 2004 (revised), "Taxes, Regulations, and the Value of U.S. and U.K. Corporations," Federal Reserve Bank of Minneapolis Staff Report 309.

[52] Mendoza, Enrique, 1997, "Terms of Trade Uncertainty and Economic Growth," Journal of Development Economics, 54, pp: 323-356.

[53] Obstfeld, M., 1994, "Risk-Taking, Global Diversification and Growth," American Economic Review, (December).

[54] Phelps, E.S., 1962, "The Accumulation of Risky Capital: A Sequential Utility Analysis," Econometrica, 30, pp: 729-743.

[55] Ramey, G. and V. Ramey, 1995, "Cross-Country Evidence on the Link Between Volatility and Growth," American Economic Review, 85, pp:1138-1151.

[56] Rebelo, S., 1991, "Long Run Policy Analysis and Long Run Growth," Journal of Political Economy, 99, 500-521.

[57] Sakellaris, P. and A. Spilimbergo, 2000, "Business Cycle and Investment in Human Capital: International Evidence on Higher Education," Carnegie-Rochester Series on Economic Policy, Vol. 52, pp: 221-256.

[58] Shell, K., 1967, "A Model of Inventive Activity and Capital Accumulation," in: K. Shell, ed., **Essays on the Theory of Optimal Economic Growth**,(MIT Press, Cambridge).

[59] Shell, K., 1973, "Inventive Activity, Industrial Organization and Economic Activity," in J. Mirrlees and N. Stern, eds., **Models of Economic Growth**, (Macmillian & Co., London, U.K)

[60] Siegler, M. V., 2001, "International Growth and Volatility in Historical Perspective," working paper, department of economics, Williams College, (November).

[61] Stokey, N.L. and R.E. Lucas (with the collaboration of E.C. Prescott), 1989, Recursive Methods in Economic Dynamics, Harvard University Press.

[62] Stokey, N. and S. Rebelo, 1995, "Growth Effects of Flat-Tax Rates," Journal of Political Economy, Vol. 103, pp: 519-550.

[63] Summers, R. and A. Heston, 1991, "The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950-1988," Quarterly Journal of Economics, 106, 2, May, pp: 327-368.

[64] Summers, R. and A. Heston, 1993, "Penn World Tables, Version 5.5," available on diskette from N.B.E.R.

[65] Turnovsky, S. J., 1995, **Methods of Macroeconomic Dynamics**, MIT Press, Cambridge, Mass, and London, Eng.

[66] Uzawa, H., 1964, "Optimal Growth in a Two Sector Model of Capital Accumulation," Review of Economic Studies, 31, pp: 1-24.

Growth with Quality-Improving Innovations: An Integrated Framework*

Philippe Aghion† and Peter Howitt‡

September 20, 2004

1 Introduction

Technological progress, the mainspring of long-run economic growth, comes from innovations that generate new products, processes and markets. Innovations in turn are the result of deliberate research and development activities that arise in the course of market competition. These Schumpeterian observations constitute the starting point of that branch of endogenous growth theory built on the metaphor of quality improvements, whose origins lie in the partial-equilibrium industrial-organization literature on patent races. Our own entry to that literature was the pre-publication version of chapter 10 of Tirole (1988).

We argued in Aghion and Howitt (1998) that by using Schumpeter's insights to develop a growth model with quality-improving innovations one can provide an integrated framework for understanding not only the macroeconomic structure of growth but also the many microeconomic issues regarding incentives, policies and institutions that interact with growth. Who gains from innovations, who loses and how much all depend on institutions and policies. By focusing on these influences in a model where entrepreneurs introduce new technologies that render previous technologies obsolete we hope to understand why those who would gain from growth prevail in some societies, while in others they are blocked by those who would lose.

In this chapter we show that the growth model with quality-improving innovations (also referred to as the "Schumpeterian" growth paradigm) is not only versatile but also simple and empirically useful. We illustrate its versatility by showing how it sheds light on such diverse issues as cross-country convergence, the effects of product-market competition on growth, and the interplay between growth and the process of institutional change. We illustrate its simplicity by building our analysis around an elementary discrete-time model. We illustrate its empirical usefulness by summarizing recent papers and studies that test

*Draft of chapter for the forthcoming Handbook of Economic Growth.

†Harvard University

‡Brown University

the microeconomic and macroeconomic implications of the framework and that address what might seem like empirically questionable aspects of the earliest prototype models in the literature.

The paper is organized as follows. Section 2 develops the basic framework. Section 3 uses it to analyze convergence and divergence patterns in the cross-country data. Section 4 analyses the interaction between growth and product market competition. Section 5 deals with the scale effect of growing population. Section 6 analyzes the interplay between institutional change and technological change, and section 7 provides some concluding remarks and suggestions for future research.

2 Basic framework

2.1 A toy version of the Aghion-Howitt model

Asked by colleagues to show them the simplest possible version of the quality-ladder model of endogenous growth which they could teach to second year undergraduate students, we came out with the following stripped-down version of Aghion-Howitt (1992).

Time is discrete, indexed by $t = 1, 2, ...$, and at each point in time there is a mass L of individuals, each endowed with one unit of skilled labor that she supplies inelastically. Each individual lives for one period and thus seeks to maximize her consumption at the end of her period.

Each period a final good is produced according to the Cobb-Douglas technology:

$$y = Ax^{\alpha}, \tag{1}$$

where x denotes the quantity of intermediate input used in final good production, and A is a productivity parameter that reflects the current quality of the intermediate good.

The intermediate good is itself produced using labor according to a simple one-for-one technology, with one unit of labor producing one unit of the current intermediate good. Thus x also denotes the amount of labor currently employed in manufacturing. But labor can also be employed in research to generate innovations.

Each innovation improves the quality of the intermediate input, from A to γA where $\gamma > 1$ measures the size of the innovation. Innovations result from research investment. More specifically, there is an innovator who, if she invests z units of labor in research, innovates with probability λz and thereby discover an improved version of the intermediate input.

The innovator enjoy monopoly power in the production of the intermediate good, but faces a competitive fringe who can produce a unit of the same intermediate good by using $\chi > 1$ units of labor instead of one. For $\chi < 1/\alpha$, this competitive fringe is binding which means that χw_t is the maximum price the innovator can charge without being driven out of the market. Her profit is thus

equal to:

$$\pi_t = (\chi - 1) w_t x_t,$$

where $w_t x_t$ is the wage bill. This monopoly rent, however, is assumed to last for one period only, after which imitation allows other individuals to produce intermediate goods of the same quality.

The model is entirely described by two equations. The first is a *labor market clearing* equation, which states that at each period total labor supply L is equal to manufacturing labor demand x plus total research labor n, that is: $L = x_t + n_t$ for all t. The second is a *research arbitrage* equation which says that in equilibrium at any date t the amount of research undertaken by the innovator must equate the marginal cost of a unit of research labor with the expected marginal benefit. The marginal cost is just the manufacturing wage w_t. The expected benefit comes from raising the probability of success by $\lambda.1 = \lambda$, in which case she earns the monopoly profit π_t involved in producing the intermediate good for the final good sector. Thus the research arbitrage equation can be expressed as:

$$w_t = \lambda \gamma \pi_t. \qquad \text{(research arbitrage)}$$

where the factor γ on the right-hand side of the equation, simply stems from the fact that an innovation multiplies wages and profits by γ.

Using the fact that the allocation of labor between research and manufacturing, remains constant in steady-state, we can drop time subscripts. Then, substituting for π_t in the research arbitrage equation, dividing through by w, and using the labor market clearing equation to substitute for x, we obtain:

$$1 = \gamma(\chi - 1)(L - n)$$

which solves for the steady-state amount of research labor, namely:

$$n = L - \frac{1}{\gamma(\chi - 1)}.$$

The equilibrium expected rate of productivity growth in steady-state, is then simply given by:

$$g = \lambda n(\gamma - 1)$$

and it therefore depends upon the characteristics of the economic environment as described by the parameters λ, γ, χ, and L. In Section 2.3 below we interpret the comparative statics of growth with respect to all these parameters, and suggest preliminary policy conclusions.

The model is extremely simple, although at the cost of making some oversimplifying assumptions. In particular, we assumed only one intermediate sector, and that labor is the only input into research.. In the next sections we relax these two assumptions. We develop a slightly more elaborated version of the quality-ladder model that we then extend in several directions to capture important aspects of the growth and development process.

2.2 A generalization

There are three kinds of goods in the economy: a general-purpose good, a large number m of different specialized intermediate inputs, and labor. Time is discrete, indexed by $t = 1, 2, ...$, and there is a mass L of individuals, each endowed with one unit of skilled labor that she supplies inelastically.[1]

The general good is produced competitively using intermediate inputs and labor, according to the production function:

$$y_t = \left(\Sigma_1^m A_{it}^{1-\alpha} x_{it}^{\alpha}\right) (L/m)^{1-\alpha} \tag{2}$$

where each x_{it} is the flow of intermediate input i used at date t, and A_{it} is a productivity variable that measures the quality of the input. The general good is used in turn for consumption, research, and producing the intermediate inputs.

The expected growth rate of any given productivity variable A_{it} is:

$$g = E\left(A_{it}/A_{i,t-1}\right) - 1 \tag{3}$$

There is no i subscript on g because, as we shall see, all sectors are *ex ante* identical and hence will have the same productivity-growth rate. Likewise there is no t subscript because, as we shall also see, the system will go immediately to a constant steady-state expected growth rate.

Productivity growth in any sector i results from innovations, which create improved versions of that intermediate input. More precisely, each innovation at t multiplies the pre-existing productivity parameter $A_{i,t-1}$ of the best available input by a factor $\gamma > 1$. Innovations in turn result from research. If N_{it} units of the general good are invested at the beginning of the period, some individual can become the new "leading-edge" producer of the intermediate input with probability μ_{it}, where:[2]

$$\mu_{it} = \lambda f\left(n_{it}\right), \qquad f' > 0, f'' < 0, f\left(0\right) = 0,$$

and $n_{it} \equiv \frac{N_{it}}{\gamma A_{i,t-1}}$ is productivity-adjusted R&D expenditure in the sector. We divide by $\gamma A_{i,t-1}$, the targeted productivity parameter, to take into account the "fishing-out" effect - on average each quality improvement is harder to bring about than the previous one.

[1] The model we present here is a simplified discrete-time version of the Aghion-Howitt (1992) model of creative destruction, which draws upon Acemoglu-Aghion-Zilibotti (2002). Grossman and Helpman (1991) presented a variant of the framework in which the x's are final consumption goods and utility is log-linear. An early attempt at developing a Schumpeterian growth model with patent races in deterministic terms was presented by Segerstrom et al (1989). Corriveau (1991) developed an elegant discrete-time model of growth through cost-reducing innovations.

[2] More precisely, $f\left(n\right) = F\left(n, k\right)$ where k is some specialized research factor in fixed supply and F is a constant-returns function. Since there is free entry in research, the equilibrium price of k adjusts so that the expected profit of an R&D firm is zero. Since this price plays no role in the analysis of growth we suppress the explicit representation of k and deal only with the decreasing-returns function f. (Of course the constant-returns assumption can be valid only over some limited range of inputs, since F is bounded above by unity.)

Assume the time period is short enough that we may ignore the possibility of more than one successful innovator in the same sector. Then:

$$A_{it} = \left\{ \begin{array}{ll} \gamma A_{i,t-1} & \text{with probability } \lambda f(n_{it}) \\ A_{i,t-1} & \text{with probability } 1 - \lambda f(n_{it}) \end{array} \right\} \tag{4}$$

According to (3) and (4) the expected productivity-growth rate in each sector can be expressed as the product of the frequency of innovations $\lambda f(n)$ and the incremental size of innovations $(\gamma - 1)$:

$$g = \lambda f(n)(\gamma - 1) \tag{5}$$

in an equilibrium where productivity-adjusted research is the same constant n in each sector. We assume moreover that the outcome of research in any one sector is statistically independent of the outcome in every other sector.

The model determines research n, and therefore the expected productivity-growth rate g, using a research arbitrage equation that equates the expected cost and benefit of research. The payoff to research in any sector i is the prospect of a monopoly rent π_{it} if the research succeeds in producing an innovation. This rent lasts for one period only, as all individuals can imitate the current technology next period. Hence the expected benefit from spending one unit on research is π_{it} times the marginal probability $\lambda f'(n) / (\gamma A_{i,t-1})$:

$$1 = \lambda f'(n)(\pi_{it} / (\gamma A_{i,t-1}))$$

To solve this equation for n we need to determine the productivity-adjusted monopoly rent π_{it}/A_{it} to a successful innovator. As before, we assume that this innovator can produce the leading-edge input at a constant marginal cost of one unit of the general good. But she faces a competitive fringe of imitators who can produce the same product at higher marginal cost χ, where $\chi \in (1, 1/\alpha)$[3] is an inverse measure of the degree of product market competition or imitation in the economy.[4]. Thus her monopoly rent is again equal to:

$$\pi_{it} = (p_{it} - 1)x_{it} = (\chi - 1)x_{it}. \tag{6}$$

A monopolist's output x_{it} will be the amount demanded by firms in the general sector when faced with the price χ; that is, the quantity such that χ equals the marginal product of the $i$the intermediate good in producing the general good:

$$\chi = \partial y_t / \partial x_{it} = \alpha (m x_{it} / A_{it} L)^{\alpha - 1} \tag{7}$$

Hence:

$$\pi_{it} = \delta(\chi) A_{it} L / m \tag{8}$$

[3] It is easily verified that if there were no fringe then the unconstrained monopolist would charge a price equal to $1/\alpha$, but (**??**) implies that at that price the fringe could profitably undercut her.

[4] If no innovation occurs then some firm will produce, but with no cost advantage over the fringe because everyone is able to produce last period's intermediate input at a constant marginal cost of unity.

where

$$\delta(\chi) \equiv (\chi - 1)(\chi/\alpha)^{\frac{1}{\alpha-1}},\ \delta'(\chi) > 0.^{5}$$

Therefore we can write the research arbitrage equation, taking into account that $\gamma A_{i,t-1} = A_{it}$ because a monopolist is someone who has just innovated, as:

$$1 = \lambda f'(n)\,\delta(\chi)\,L/m \tag{9}$$

which we assume in this section has a positive solution.

The expected productivity growth rate is determined by substituting the solution of (9) into the growth equation (5). In the special case where the research-productivity function f takes the simple form:

$$f(n) = \sqrt{2n},$$

we have:

$$g = \lambda^2 \delta(\chi)\,(L/m)\,(\gamma - 1) \tag{10}$$

As it turns out, g is not only the expected growth rate of each sector's productivity parameter but also the approximate growth rate of the economy's per-capita GDP. This is because per-capita GDP is approximately proportional to the unweighted average of the sector-specific productivity parameters:[6]

$$A_t = \frac{1}{m}\sum_{i=1}^{m} A_{it}.$$

[5] To see that $\delta' > 0$ note that:

$$\chi \frac{d\ln(\delta(x))}{d\chi} = \frac{\chi}{\chi - 1} - \frac{1}{1-\alpha} > 0$$

where the last inequality follows from (**??**).

[6] To see this, note that GDP equals the sum of value added in the general sector and in the monopolized intermediate sectors. (There is no value added in the competitive intermediate sectors because their output is priced at the cost of the intermediate inputs. Also, we follow the standard national income accounting practice of ignoring the output (patents) of the research sector.) According to (7) the output of each monopolized sector ($i \in M_t$) at t is:

$$x_{it} = (\alpha/\chi)^{\frac{1}{1-\alpha}}\,(L/m)\,A_{it} \qquad (i \in M_t)$$

The output of each competitive sector ($i \in C_t$) at t is the amount demanded when its price is unity; setting $\chi = 1$ in (7) yields:

$$x_{it} = (\alpha)^{\frac{1}{1-\alpha}}\,(L/m)\,A_{it} \qquad (i \in C_t)$$

Substituting these into (2) and rearranging yields the following expression for per-capita GDP:

$$y_t/L = (\alpha/\chi)^{\frac{1}{1-\alpha}}\left(\frac{\#M_t}{m}\right)\left(\frac{1}{\#M_t}\sum_{i\in M_t} A_{it}\right) + \alpha^{\frac{1}{1-\alpha}}\left(\frac{\#C_t}{m}\right)\left(\frac{1}{\#C_t}\sum_{i\in C_t} A_{it}\right)$$

where $\#M_t$ is the number of monopolized sectors and $\#C_t$ the number of competitive sectors. By the law of large numbers, the fraction of sectors $\#M_t/m$ monopolized, i.e. the fraction in which there was an innovation at t, is approximately the probability of success in research in each sector: $\mu = \lambda f(n)$, and the fraction of sectors $\#C_t/m$ that are competitive is approximately $1 - \mu$. The average productivity parameter among monopolized sectors $\left(\frac{1}{\#M_t}\sum_{i\in M_t} A_{it}\right)$ is just γ times the average productivity parameter of those sectors last period; since innovations are spread randomly across sectors this is approximately γ times the average across all sectors last period: γA_{t-1}. Likewise the average productivity parameter

Since (a) all sectors have an expected growth rate of g, (b) the sectoral growth rates are statistically independent of each other and (c) there is a large number of them, therefore the law of large numbers implies that the average grows at approximately the same rate g as each component.

2.3 Alternative formulations

There are many other ways of formulating the basic model. We note two of them here for future reference. In the first one, as in the above toy model, the general good is used only for consumption, while skilled labor is the only factor used in producing intermediate products and research. The general good is produced by the intermediate inputs in combination with a specialized factor (for example unskilled labor) available in fixed supply. In this formulation, the growth equation is the same as (5) above, but with n being interpreted as the amount of skilled labor allocated to R&D. This version will be spelled out in somewhat more detail in section 5 below.

The other popular version is one with intersectoral spillovers, in which each innovation produces a new intermediate product in that sector embodying the maximum $\overline{A}_{t-1}$ of all productivity parameters of the last period, across all sectors, times some factor γ that depends on the flow of innovations in the whole economy. The idea here is that if a sector has been unlucky for a long time, while the rest of the economy has progressed, the technological progress elsewhere spills over into the innovation in this sector, resulting in a larger innovation than if the innovation had occurred many years ago. The model in section 3 below is a variant of this version.

2.4 Comparative statics on growth

Equation (10) delivers several comparative-statics results, each with important policy implications on how to "manage" the growth process:

1. Growth increases with the productivity of innovations λ and with the

---

across competitive sectors is the average across sectors that did not innovate, which is approximately A_{t-1}. Making these substitutions into the above expression for per-capita GDP yields:

$$y_t/L \simeq \left((\alpha/\chi)^{\frac{1}{1-\alpha}}\mu\gamma + \alpha^{\frac{1}{1-\alpha}}(1-\mu)\right)A_{t-1} \equiv \zeta A_{t-1}$$

Since labor is paid its marginal product in the general sector, the wage rate is:

$$w_t = \partial y_t/\partial L = (1-\alpha)\, y_t/L \simeq (1-\alpha)\,\zeta A_{t-1}$$

which is also per-capita value-added in the general sector. By similar reasoning, (8) implies that per-capita value added in monopolized intermediate sectors is:

$$(1/L)\sum_{i\in M_t}\delta(\chi)\,A_{it}L/m = \delta(\chi)\left(\frac{\#M_t}{m}\right)\left(\frac{1}{\#M_t}\sum_{i\in M_t}A_{it}\right) \simeq \delta(\chi)\,\mu\gamma A_{t-1}.$$

Therefore each component of per-capita GDP is approximately proportional to A_{t-1}. Since A_t grows at approximately the constant rate g thefore per-capita GDP is approximately proportional to A_t.

supply of skilled labor L: both results point to the importance of education, and particularly higher education, as a growth-enhancing factor. Countries that invest more in higher education will achieve a higher productivity of research activities and also reduce the opportunity cost of R&D by increasing the aggregate supply of skilled labor. An increase in the size of population should also bring about an increase in growth by raising L. This "scale effect" has been challenged in the literature and will be discussed in section 5 below.

2. Growth increases with the size of innovations, as measured by γ. This result points to the existence of a wedge between private and social innovation incentives. That is, a decrease in size would reduce the cost of innovation in proportion to the expected rents; the research arbitrage equation (9) shows that these two effects cancel each other, leaving the equilibrium level of R&D independent of size. However, equation (10) shows that the social benefit from R&D, in the form of enhanced growth, is proportional not to γ but to the "incremental size" $\gamma - 1$. When γ is close to one it is not socially optimal to spend spend as much on R&D as when γ is very large, because there is little social benefit; yet a laissez-faire equilibrium would result in the same level of R&D in both cases.

3. Growth is decreasing with the degree of product market competition and/or with the degree of imitation as measured inversely by χ. Thus patent protection (or, more generally, better protection of intellectual property rights), will enhance growth by increasing χ and therefore increasing the potential rewards from innovation. However, pro-competition policies will tend to discourage innovation and growth by reducing χ and thereby forcing incumbent innovators to charge a lower limit price. Existing historical evidence supports the view that property rights protection is important for sustained long-run growth; however the prediction that competition should be unambiguously bad for innovations and growth is questioned by all recent empirical studies, starting with the work of Nickell (1996) and Blundell et al (1999). In Section 4 we shall argue that the Schumpeterian framework outlined in this section can be extended so as to reconcile theory and evidence on the effects of entry and competition on innovations, and that it also generates novel predictions regarding these effects which are borne out by empirical tests.

3 Linking growth to development: convergence clubs

With its emphasis on institutions, the Schumpeterian growth paradigm is not restricted to dealing with advanced countries that perform leading-edge R&D. It can also shed light on why some countries that were initially poor have managed to grow faster than industrialized countries, whereas others have continued to

fall further behind.

The history of cross-country income differences exhibits mixed patterns of convergence and divergence. The most striking pattern over the long run is the "great divergence" - the dramatic widening of the distribution that has taken place since the early 19th Century. Pritchett (1997) estimates that the proportional gap in living standards between the richest and poorest countries grew more than five-fold from 1870 to 1990, and according to the tables in Maddison (2001) the proportional gap between the richest group of countries and the poorest[7] grew from 3 in 1820 to 19 in 1998. But over the second half of the twentieth century this widening seems to have stopped, at least among a large group of nations. In particular, the results of Barro and Sala-i-Martin (1992), Mankiw, Romer and Weil (1992) and Evans (1996) seem to imply that most countries are converging to parallel growth paths.

However, the recent pattern of convergence is not universal. In particular, the gap between the leading countries as a whole and the very poorest countries as a whole has continued to widen. The proportional gap in per-capita income between Mayer-Foulkes's (2002) richest and poorest convergence groups grew by a factor of 2.6 between 1960 and 1995, and the proportional gap between Maddison's richest and poorest groups grew by a factor of 1.75 between 1950 and 1998. Thus as various authors[8] have observed, the history of income differences since the mid 20th Century has been one of "club-convergence"; that is, all rich and most middle-income countries seem to belong to one group, or "convergence club", with the same long-run growth rate, whereas all other countries seem to have diverse long-run growth rates, all strictly less than that of the convergence club.

The explanation we develop in this section for club convergence follows Howitt (2000), who took the cross-sectoral-spillovers variant of the closed-economy model described in section 2.3 above and allowed the spillovers to cross international as well as intersectoral borders. This international spillover, or "technology transfer", allows a backward sector in one country to catch up with the current technological frontier whenever it innovates. Because of technology transfer, the further behind the frontier a country is initially, the bigger the average size of its innovations, and therefore the higher its growth rate for a given frequency of innovations. As long as the country continues to innovate at some positive rate, no matter how small, it will eventually grow at the same rate as the leading countries. (Otherwise the gap would continue to rise and therefore the country's growth rate would continue to rise.) However, countries with poor macroeconomic conditions, legal environment, education system or credit markets will not innovate in equilibrium and therefore they will not benefit from technology transfer, but will instead stagnate.

This model reconciles Schumpeterian theory with the evidence to the effect

[7] The richest group was Western Europe in 1820 and the "Euopean Offshoots" (Australia, Canada, New Zealand and the United States) in 1998. The poorest group was Africa in both years.

[8] Baumol (1986), Durlauf and Johnson (1995), Quah (1993, 1997) and Mayer-Foulkes (2002, 2003).

that all but the poorest countries have parallel long-run growth paths. It implies that the growth rate of any one country depends not on local conditions but on global conditions that impinge on world-wide innovation rates. The same parameters which were shown in section 2.4 above to determine a closed economy's productivity-growth rate will now determine that country's relative productivity *level*. What emerges from this exercise is therefore not just a theory of club convergence but also a theory of the world's growth rate and of the cross-country distribution of productivity.

Before we develop the model we need to address the question of how our framework, in which growth depends on research and development, can be applied to the poorest countries of the world, in which, according to OECD statistics, almost no formal R&D takes place. The key to our answer is that because technological knowledge is often tacit and circumstantially specific,[9] foreign technologies cannot simply be copied and transplanted to another country no cost. Instead, technology transfer requires the receiving country to invest resources in order to master foreign technologies and adapt them to the local environment. Although these investments may not fit the conventional definition of R&D, they play the same role as R&D in an innovation-based growth model; that is, they use resources, including skilled labor with valuable alternative uses, they generate new technological possibilities where they are conducted, and they build on previous knowledge.[10] While it may be the case that implementing a foreign technology is somewhat easier than inventing an entirely new one, this is a difference in degree, not in kind. In the interest of simplicity our theory ignores that difference in degree and treats the implementation and adaptation activities undertaken by countries far behind the frontier as being analytically the same as the research and development activities undertaken by countries on or near the technological frontier. For all countries we assign to R&D the role that Nelson and Phelps (1966) assumed was played by human capital, namely that of determining the country's "absorptive capacity".[11]

3.1 A model of technology transfer

Consider one country in a world of h different countries. This country looks just like the ones described in the basic model above, except that whenever an innovation takes place in any given sector the productivity parameter attached to the new product will be an improvement over the pre-existing global leading-edge technology. That is, let $\overline{A}_{t-1}$ be the maximum productivity parameter over all countries in the sector at the end of period $t-1$; in other words the "frontier" productivity at $t-1$. Then an innovation at date t will result in a new version

[9] See Arrow (1969) and Evenson and Westphal (1995).

[10] Cohen and Levinthal (1989) and Griffith, Redding and Van Reenen (2001) have also argued that R&D by the receiving country is a necessary input to technology transfer.

[11] Grossman and Helpman (1991) and Barro and Sala-i-Martin (1997) also model technology transfer as taking place through a costly investment process, which they portray as imitation; but in these models technology transfer always leads to convergence in growth rates except in special cases studied by Grossman and Helpman where technology transfer is inactive in the long run.

of that intermediate sector whose productivity parameter is $A_t = \gamma \overline{A}_{t-1}$, which can be implemented by the innovator in this country, and which becomes the new global frontier in that sector. The frontier parameter will also be raised by the factor γ if an innovation occurs in that sector in any other country.

Therefore domestic productivity in the sector evolves according to:

$$\ln A_t = \left\{ \begin{array}{ll} \ln \overline{A}_{t-1} + \ln \gamma = \ln \overline{A}_t & \text{with probability } \mu \\ \ln A_{t-1} & \text{with probability } 1-\mu \end{array} \right\}$$

where μ is the country's innovation rate:

$$\mu = \lambda f(n)$$

and the productivity-adjusted research n is defined as:

$$n \equiv N_t / \left(\gamma \overline{A}_{t-1}\right)$$

since the targeted productivity parameter is now $\gamma \overline{A}_{t-1}$.

Meanwhile, the global frontier advances by the factor γ with every innovation anywhere in the world.[12] Therefore:

$$\ln \overline{A}_t = \left\{ \begin{array}{ll} \ln \overline{A}_{t-1} + \ln \gamma & \text{with probability } \overline{\mu} \\ \ln \overline{A}_{t-1} & \text{with probability } 1-\overline{\mu} \end{array} \right\} \tag{11}$$

where

$$\overline{\mu} = \Sigma_1^h \lambda^j f\left(n^j\right)$$

is the global innovation rate.[13] It follows from (11) that the long-run average growth rate of the frontier, measured as a difference in logs, is:[14]

$$\overline{g} = \overline{\mu} \ln \gamma \tag{12}$$

Assume there is no international trade in intermediates or general goods.[15] Then the costs and benefits of R&D are the same as in the previous model,

[12] Again we are assuming a time period small enough to ignore the possibility of simultaneous innovations in the same sector.

[13] A simpler version of the model would just have the frontier productivity grow at an exogenous rate $\overline{g}$. The model in this section has the advantage of delivering both an endogenous rate for productivity growth at the frontier and club convergence towards that frontier.

[14] The growth rate (12) expressed as a log difference is approximately the same as the rate (5) of the previous section which was expressed as a proportional increment, because the first-order Taylor-series approximation to $\ln\gamma$ at $\gamma = 1$ is $(\gamma - 1)$. We switch between these two definitions depending on which is more convenient in a given context.

[15] This is not to say that international trade is umimportant for technology transfer. On the contrary, Coe and Helpman (1995), Coe Helpman and Hoffmaister (1996), Eaton and Kortum (1996) and Savvides and Zachariadis (2004) all provide strong evidence to the effect that internation trade plays an important role in the international diffusion of technological progress. For a recent summary of this and other empirical work, see Keller (2002). Eaton and Kortum (2001) provide a simple "semi-endogenous" (see section 5 below) growth model in which endogenous innovation interacts with technology transfer and international trade in goods; in their model all countries converge to the same long-run growth rate.

except that the domestic productivity parameter A_t may differ from the global parameter $\overline{A}_t$ that research aims to improve upon. Each innovation will now change log productivity by:

$$\ln \overline{A}_{t-1} + \ln \gamma - \ln A_{t-1} = \ln \gamma + d_{t-1}$$

where

$$d_{t-1} \equiv \ln \left(\overline{A}_{t-1} / A_{t-1} \right)$$

is a measure of "distance to the frontier." As Gerschenkron (1952) argued when discussing the "advantage of backwardness," the greater the distance the larger the innovation. The average growth rate will again be the expected frequency of innovations times size:

$$g_t = \mu \left(\ln \gamma + d_{t-1} \right) \tag{13}$$

which is also larger the greater the distance to the frontier.

The distance variable d_t evolves according to:

$$d_t = \left\{ \begin{array}{cc} d_{t-1} & \text{with probability } 1 - \overline{\mu} \\ \ln \gamma + d_{t-1} & \text{with probability } \overline{\mu} - \mu \\ 0 & \text{with probability } \mu \end{array} \right\}$$

That is, with probability $1 - \overline{\mu}$ there is no innovation in the sector either globally or in this country, so both domestic productivity and frontier productivity remain unchanged; with probability $\overline{\mu} - \mu$ an innovation will occur in this sector but in some other country, in which case domestic productivity remains the same but the proportional gap grows by the factor γ; and with probability μ an innovation will occur in this sector in this country, in which case the country moves up to the frontier, reducing the gap to zero.

It follows that the expected distance $\widehat{d}_t$ evolves according to:

$$\widehat{d}_t = (1 - \mu) \, \widehat{d}_{t-1} + (\overline{\mu} - \mu) \ln \gamma.$$

If $\mu > 0$ this is a stable difference equation with a unique rest point. That is, as long as the country continues to perform R&D at a positive constant intensity n its distance to the frontier will stabilize, meaning that its productivity growth rate will converge to that of the global frontier. But if $\mu = 0$ the difference equation has no stable rest point and $\widehat{d}_t$ diverges to infinity. That is, if the country stops innovating it will have a long-run productivity growth rate of zero because innovation is a necessary condition for the country to benefit from technology transfer.

More formally, the country's long-run expected distance d^* is given by:

$$d^* = \left\{ \begin{array}{ll} (\overline{\mu}/\mu - 1) \ln \gamma & \text{if } \mu > 0 \\ \infty & \text{if } \mu = 0 \end{array} \right\} \tag{14}$$

and its long-run expected growth rate g^*, according to (12), (13) and (14) is:

$$g^* = \left\{ \begin{array}{ll} \mu \left(\ln \gamma + d^* \right) = \overline{g} & \text{if } \mu > 0 \\ 0 & \text{if } \mu = 0 \end{array} \right\}$$

Each country's innovation rate μ is determined according to the same principles as before. In particular, it will be equal $\lambda f(n)$ where n is determined by the research-arbitrage equation (9) above, provided that a positive solution to (9) exists. For example if the research-productivity function f satisfies the Inada-like condition: $f'(0) = \infty$, as in the example used above to derive the growth equation (10), then there will always exist a positive solution to (9), so all countries will converge to the frontier growth rate.

But suppose, on the contrary, that this Inada-like condition does not hold, that instead: $f'(0) < \infty$. Then the research-arbitrage condition (9) must be replaced by the more general Kuhn-Tucker conditions:

$$1 \geq \lambda f'(n)\,\delta(\chi)\,L/m \qquad \text{with } n = 0 \text{ if the inequality is strict.} \tag{15}$$

That is, for an interior solution the expected marginal cost and benefit must be equal, but the only equilibrium will be one with zero R&D if at that point the expected marginal benefit does not exceed the cost. It follows that the country will perform positive R&D if:

$$\lambda\delta(\chi)\,L/m > 1/f'(0), \tag{16}$$

but if condition (16) fails then there will be no research: $n = 0$ and hence no innovations: $\mu = 0$ and no growth: $g = 0$.

This means that countries will fall into two groups, corresponding to two convergence clubs:

1. Countries with highly productive R&D, as measured by λ, or good educational systems as measured by high λ or high L, or good property right protection as measured by a high χ, will satisfy condition (16), and hence will grow asymptotically at the frontier growth rate $\overline{g}$.

2. Countries with low R&D productivity, poor educational systems and low property right protection will fail condition (16) and will not grow at all. The gap d_t separating them from the frontier will grow forever at the rate $\overline{g}$.

3.2 World growth and distribution

Since the world growth rate $\overline{g}$ given by (12) depends on each country's innovation frequency $\mu^j = \lambda^j f\left(n^j\right)$, therefore world growth depends on the value for each country of all the factors described in section 2.4 above that determine μ^j. Thus any improvement in R&D productivity, education or property rights anywhere in the innovating world will raise the growth rate of productivity in all but the stagnating countries.

Moreover, the cross-country distribution of productivity is determined by these same variables. For according to (14) each country's long-run relative distance to the frontier depends uniquely on its own innovation frequency $\mu = \lambda f(n)$. Two countries in which the determinants of innovation analyzed in section 2.4 are the same will lie the same distance from the frontier in the long

run and hence will have the same productivity in the long run. Countries with more productive R&D, better educational systems and stronger property right protection will have higher productivity.

3.3 The role of financial development in convergence

The framework can be further developed by assuming that while the size of innovations increases with the distance to the technological frontier (due to technology transfer), the frequency of innovations depends upon the ratio between the distance to the technological frontier and the current stock of skilled workers. This enriched framework (see Howitt and Mayer-Foulkes, 2002) can explain not only why some countries converge while other countries stagnate but also why different countries may display positive yet divergent growth patterns in the long-run. Benhabib and Spiegel (2002) develop a similar account of divergence and show the importance of human capital in the process. The rest of this section presents a summary of the related model of Aghion, Mayer-Foulkes and Howitt (2004) (AMH) and discusses their empirical results showing the importance of financial development in the convergence process.

Suppose that the world is as portrayed in the previous sections, but that research aimed at making an innovation in t must be done at period $t-1$. If we assume perfectly functioning financial markets then nothing much happens to the model except that the research arbitrage condition (9) has a discount factor β on the right-hand side to reflect the fact that the expected returns to R&D occur one period later than the expenditure.[16] But when credit markets are imperfect, AMH show that an entrepreneur may face a borrowing constraint that limits her investment to a fixed multiple of her accumulated net wealth. In their model the multiple comes from the possibility that the borrower can, at a cost that is proportional to the size of her investment, decide to defraud her creditors by making arrangements to hide the proceeds of the R&D project in the event of success.[17] They also assume a two-period overlapping-generations structure in which the accumulated net wealth of an entrepreneur is her current wage income, and in which there is just one entrepreneur per sector in each country. This means that the further behind the frontier the country falls the less will any entrepreneur be able to invest in R&D relative to what is needed to maintain any given frequency of innovation. What happens in the long run to the country's growth rate depends upon the interaction between this disadvantage of backwardness, which reduces the frequency of innovations, and the above-described advantage of backwardness, which increases the size of innovations. The lower the cost of defrauding a creditor the more likely it is that the disadvantage of backwardness will be the dominant force, preventing the country from converging to the frontier growth rate even in the long run. Generally speaking, the greater the degree of financial development of a country the

[16] For simplicity we suppose that everyone has linear intertemporal preferences with a constant discount factor β.

[17] The "credit multipler" assumed here is much like that of Bernanke and Gertler (1989), as modified by Aghion, Banerjee and Piketty (1999).

more effective are the institutions and laws that make it difficult to defraud a creditor. Hence the link between financial development and the likelihood that a country will converge to the frontier growth rate.

The following simplified account of AHM shows in more detail how this link between financial development and convergence works. Suppose that entrepreneurs have no source of income other than what they can earn from innovating. Then they must borrow the entire cost of any R&D project. Because there are constant returns to the R&D technology,[18] therefore in equilibrium that cost will equal the expected benefit, discounted back to today:

$$\mu\beta\pi_t = \left(\lambda N_t / \left(\gamma \overline{A}_t\right)\right) \beta\delta\left(\chi\right) \gamma \overline{A}_t = \lambda\beta\delta\left(\chi\right) N_t$$

This is also the expected discounted benefit to a borrower from paying a cost cN_t today that would enable her to default in the event that the R&D project is successful. (There is no benefit if the project fails to produce an innovation because in that case the entrepreneur cannot pay anything to the creditor even if she has decided to be honest and therefore has not paid the cost cN_t). The entrepreneur will choose to he honest if the cost at least as great as the benefit; that is, if:

$$c \geq \lambda\beta\delta\left(\chi\right). \tag{17}$$

Otherwise she will default on any loan.

Suppose that $\beta\lambda\delta\left(\chi\right) L/m > 1/f'\left(0\right)$. This is the condition (16) above for positive growth, modified to take discounting into account. It follows that in any country where the incentive-compatibility constraint (17) holds then innovation will proceed as described in the previous section, and the country will converge to the frontier growth rate. But in any country where the cost of defrauding a creditor is less than the right-hand side of (17) no R&D will take place because creditors would rationally expect to be defrauded of any possible return from lending to an entrepreneur. Therefore convergence to the frontier growth rate will occur only in countries with a level of financial development that is high enough to put the cost of fraud at or above the limit imposed by (17).

AHM test this effect of financial development on convergence by running the following cross-country growth regression:

$$g_i - g_1 = \beta_0 + \beta_f F_i + \beta_y \cdot \left(y_i - y_1\right) + \beta_{fy} \cdot F_i \cdot \left(y_i - y_1\right) + \beta_x X_i + \varepsilon_i \tag{18}$$

where g_i denotes the average growth rate of per-capita GDP in country i over the period 1960 - 1995, F_i the country's average level of financial development, y_i the initial (1960) log of per-capita GDP, X_i a set of other regressors and ε_i a disturbance term with mean zero. Country 1 is the technology leader, which they take to be the United States.

Define $\widehat{y}_i \equiv y_i - y_1$, country i's initial relative per-capita GDP. Under the assumption that $\beta_y + \beta_{fy} F_i \neq 0$ we can rewrite (18) as:

$$g_i - g_1 = \lambda_i \cdot \left(\widehat{y}_i - \widehat{y}_i^*\right)$$

[18] See footnote 2 above.

where the steady-state value $\widehat{y}_i^*$ is defined by setting the RHS of (18) to zero:

$$\widehat{y}_i^* = -\frac{\beta_0 + \beta_f F_i + \beta_x X_i + \varepsilon_i}{\beta_y + \beta_{fy} F_i} \tag{19}$$

and λ_i is a country-specific convergence parameter:

$$\lambda_i = \beta_y + \beta_{fy} F_i \tag{20}$$

that depends on financial development.

A country can converge to the frontier growth rate if and only if the growth rate of its relative per-capita GDP depends negatively on the initial value $\widehat{y}_i$; that is if and only if the convergence parameter λ_i is negative. Thus the likelihood of convergence will increase with financial development, as implied by the above theory, if and only if:

$$\beta_{fy} < 0. \tag{21}$$

The results of running this regression using a sample of 71 countries are shown in Table 1, which indicates that the interaction coefficient β_{fy} is indeed significantly negative for a variety of different measures of financial development and a variety of different conditioning sets X. The estimation is by instrumental variables, using a country's legal origins, and its legal origins[19] interacted with the initial GDP gap $(y_i - y_1)$ as instruments for F_i and $F_i (y_i - y_1)$. The data, estimation methods and choice of conditioning sets X are all take directly from Levine, Loayza and Beck (2000) who found a strongly positive and robust effect of financial intermediation on short-run growth in a regression identical to (18) but without the crucial interaction term $F_i (y_i - y_1)$ that allows convergence to depend upon the level of financial development.

TABLE 1 HERE

AHM shown that the results of Table 1 are surprisingly robust to different estimation techniques, to discarding outliers, and to including possible interaction effects between the initial GDP gap and other right-hand-side variables.

3.4 Concluding remark

Thus we see how Schumpeterian growth theory and the quality improvement model can naturally explain club convergence patterns, the so-called twin peaks pointed out by Quah (1996). The Schumpeterian growth framework can deliver an explanation for cross-country differences in growth rates and/or in convergence patterns based upon *institutional considerations*. No one can deny that such considerations are close to what development economists have been concerned with. However, some may argue that the quality improvement paradigm,

[19] See LaPorta et al. (1998) for a detailed explanation of legal origins and its relevance as an instrument for financial development.

and new growth theories in general, remain of little help for development policy, that they merely formalize platitudes regarding the growth-enhancing nature of good property right protection, sound education systems, stable macroeconomy, without regard to specifics such as a country's current stage of development. In Section 4 and 6 below we will argue on the contrary that the Schumpeterian growth paradigm can be used to understand (i) why liberalization policies (in particular an increase in product market competition) should affect productivity growth differently in sectors or countries at different stages of technological development as measured by the distance variable d; and (ii) why the organizations or institutions that maximize growth, or that are actually chosen by societies, also vary with distance to the frontier.

4 Linking growth to IO: innovate to escape competition

One particularly unappealing feature of the basic Schumpeterian model outlined in Section 2 is the prediction that product market competition is unambiguously detrimental to growth because it reduces the monopoly rents that reward successful innovators and thereby discourages R&D investments. Not only does this prediction contradict a common wisdom that goes back to Adam Smith, but it has also been shown to be (partly) counterfactual (e.g by Geroski (1994), Nickell (1996), and Blundell et al (1999))[20].

However, as we argue in this section, a simple modification reconciles the Schumpeterian paradigm with the evidence on product market competition and innovation, and also generates new empirical predictions that can be tested with firm- and industry-level data. In this respect the paradigm can meet the challenge of seriously putting IO into growth theory. The theory developed in this section is based on Aghion-Harris-Vickers (1997) and Aghion-Harris-Howitt-Vickers (2001), but cast in the discrete-time framework introduced above.

We start by considering an isolated country in a variant of the technology-transfer model of the previous section. This variant allows technology spillovers to occur across sectors as well as across national borders. Thus there is a global technological frontier that is common to all sectors, and which is drawn on by all innovations. The model takes as given the growth rate of this global frontier, so that the frontier $\overline{A}_t$ at the end of period t obeys:

$$\overline{A}_t = \gamma \overline{A}_{t-1},$$

where $\gamma > 1$.

In each country, the general good is produced using the same kind of technology as in the previous sections, but here for simplicity we assume a continuum

[20] We refer the reader to the second part of this section where we confront theory and empirics on the relationship between competition/entry and innovation/productivity growth.

of intermediate inputs and we normalize the labor supply at $L = 1$, so that:

$$y_t = \int_0^1 A_{it}^{1-\alpha} x_{it}^{\alpha} di,$$

where, in each sector i, only one firm produces intermediate input i using general good as capital according to a one-for-one technology.

In each sector, the incumbent firm faces a competitive fringe of firms that can produce the same kind of intermediate good, although at a higher unit cost. More specifically, we assume that at the end of period t, at unit cost χ, where we assume $1 < \chi < 1/\alpha < \gamma\chi$, a competitive fringe of firms can produce one unit of intermediate input i of a quality equal to $\min(A_{it}, \overline{A}_{t-1})$, where A_{it} is the productivity level achieved in sector i after innovation has had the opportunity to occur in sector i within period t.

In each period t, there are three types of sectors, which we refer to as type-j sectors, with $j \in \{0, 1, 2\}$. A type-j sector starts up at the beginning of period t with productivity $A_{j,t-1} = \overline{A}_{t-1-j}$, that is, j steps behind the current frontier $\overline{A}_{t-1}$. The profit flow of an incumbent firm in any sector at the end of period t, will depend upon the technological position of that firm with regard to the technological frontier at the end of the period.

Between the beginning and the end of the current period t, the incumbent firm in any sector i has the possibility of innovating with positive probability. Innovations occur step-by-step: in any sector an innovation moves productivity upward by the same factor γ. Incumbent firms can affect the probability of an innovation by investing more in R&D at the beginning of the period. Namely, by investing the quadratic R&D effort $\frac{1}{2}\gamma A_{i,t-1}\mu^2$ incumbent an firm i in a type-0 or type-1 sector, innovates with probability μ.[21] However, innovation is assumed to be automatic in type-2 sectors, which in turn reflects a knowledge externality from more advanced sectors which limits the maximum distance of any sector to the technological frontier.

Now, consider the R&D incentives of incumbent firms in the different types of sectors at the beginning of period t. Firms in type-2 sectors have no incentive to invest in R&D since innovation is automatic in such sectors. Thus

$$\mu_2 = 0,$$

where μ_j is the equilibrium R&D choice in sector j.

Firms in type-1 sectors, that start one step behind the current frontier at $A_{i,t-1} = \overline{A}_{t-2}$ at the beginning of period t, end up with productivity $A_t = \overline{A}_{t-1}$ if they successfully innovate, and with productivity $A_t = \overline{A}_{t-2}$ otherwise. In either case, the competitive fringe can produce intermediate goods of the same quality but at cost χ instead of 1, which in turn, as in section 2 above, the

[21] We thus depart slightly from our formulation in the previous sections: here we take the probability of innovation, not the R&D effort, as the optimization variable. However the two formulations are equivalent: that the innovation probability $f(n) = \mu$ is a concave function of the effort n, is equivalent to saying that the effort is a convex function of the probability.

equilibrium profit is equal to:[22]

$$\pi_t = A_t \delta(\chi),$$

with

$$\delta(\chi) = (\chi - 1)(\chi/\alpha)^{\frac{1}{\alpha-1}}.$$

Thus the net rent from innovating for a type-1 firm is equal to

$$(\overline{A}_{t-1} - \overline{A}_{t-2})\delta(\chi)$$

and therefore a type-1 firm will choose its R&D effort to solve:

$$\max_{\mu}\{(\overline{A}_{t-1} - \overline{A}_{t-2})\delta(\chi)\mu - \frac{1}{2}\gamma\overline{A}_{t-2}\mu^2\},$$

which yields

$$\mu_1 = (1 - \frac{1}{\gamma})\delta(\chi).$$

In particular an increase in product market competition, measured as an reduction in the unit cost χ of the competitive fringe, will reduce the innovation incentives of a type-1 firm. This we refer to as the *Schumpeterian effect* of product market competition: competition reduces innovation incentives and therefore productivity growth by reducing the rents from innovations of type-1 firms that start below the technological frontier. This is the dominant effect, both in IO models of product differentiation and entry, and in basic endogenous growth models as the one analyzed in the previous sections. Note that type-1 firms cannot escape the fringe by innovating: whether they innovate or not, these firms face competitors that can produce the same quality as theirs at cost χ. As we shall now see, things become different in the case of type-0 firms.

Firms in type-0 sectors, that start at the current frontier, end up with productivity $\overline{A}_t$ if they innovate, and stay with their initial productivity $\overline{A}_{t-1}$ if they do not. But the competitive fringe can never get beyond producing quality $\overline{A}_{t-1}$. Thus, by innovating, a type-0 incumbent firm produces an intermediate good which is γ times better than the competing good the fringe could produce, and at unit cost 1 instead of χ for the fringe. Our assumption $\frac{1}{\alpha} < \gamma\chi$ then implies that competition by the fringe is no longer a binding constraint for an innovating incumbent, so that its equilibrium profit post-innovation, will simply be the profit of an unconstrained monopolist, namely:

$$\pi_t = \overline{A}_t \delta(1/\alpha).$$

On the other hand, a type-0 firm that does not innovate, will keep its productivity equal to $\overline{A}_{t-1}$. Since the competitive fringe can produce up to this quality level at cost χ, the equilibrium profit of a type-0 firm that does not innovate, is equal to

$$\pi_t = \overline{A}_{t-1}\delta(\chi).$$

[22] Imitation does not destroy the rents of non-innovating firms. We assume nevertheless that the firm ignores any continuation value in its R&D decision.

A type-0 firm will then choose its R&D effort to:

$$\max_{\mu}\{[\overline{A}_t\delta(1/\alpha) - \overline{A}_{t-1}\delta(\chi)]\mu - \frac{1}{2}\gamma\overline{A}_{t-1}\mu^2\},$$

so that in equilibrium

$$\mu_0 = \delta(1/\alpha) - \frac{1}{\gamma}\delta(\chi).$$

In particular an increase in product market competition, i.e a reduction in χ, will now have a fostering effect on R&D and innovation. This, we refer to as the *escape competition effect:* competition reduces pre-innovation rents of type-0 incumbent firms, but not their post-innovation rents since by innovating these firms have escaped the fringe. This in turn induces those firms to innovate in order to escape competition with the fringe.

4.1 Composition effect and the inverted-U relationship between competition and innovation

We have just seen that product market competition tends to have opposite effects on frontier and lagging sectors, fostering innovation by the former and discouraging innovation by the latter. In this section we consider the impact of competition on the steady-state aggregate innovation intensity

$$I = q_0\mu_0 + q_1\mu_1 \tag{22}$$

where q_j is the steady-state fraction of type-j sectors (recall that type-2 sectors do not perform R&D).

To get a non-trivial steady-state fraction of type-0 firms, we need that the net flows out of state 0 (which corresponds to type-0 firms that fail to innovate in the current period), be compensated by a net flow into state 0. We simply postulate such a flow into state 0, by assuming that at the end of any period t, with exogenous probability ε entry at the new frontier, that is by a type-0 firm with productivity level $\overline{A}_t$, occurs in a type-2 sector after the incumbent firm has produced. We then have the following flow equations describing the net flows into and out of states 0, 1 and 2:

$$\begin{aligned} q_2\varepsilon &= q_0(1-\mu_0); \\ q_0(1-\mu_0) &= q_1(1-\mu_1); \\ q_1(1-\mu_1) &= q_2\varepsilon; \end{aligned}$$

in which the left hand sides represents the steady-state expected flow of sectors that move into a state j and the right hand sides represent the expected outflow from the same state, for $j = 0, 1,$ and 2. This, together with the identity:

$$q_0 + q_1 + q_2 = 1,$$

implies that:

$$I = 1 - q_2(1 + 2\varepsilon),$$

where

$$q_2 = \frac{1}{1 + \frac{\varepsilon}{1-\mu_0} + \frac{\varepsilon}{1-\mu_1}}.$$

In particular, one can see that the overall effect of increased product market competition on I is ambiguous since it produces opposite effects on innovation probabilities in type-0 and type-1 sectors (i.e on μ_0 and μ_1). In fact, one can say more than that, and show that: (i) the Schumpeterian effect always dominates for γ sufficiently large; (ii) the escape competition effect always dominates for γ sufficiently close to one; (iii) for intermediate values of γ, the escape competition effect dominates when competition is initially low (with χ close to $1/\alpha$) whereas the Schumpeterian effect dominates when competition is initially high (with χ close to one). In this latter case, the relationship between competition and innovation is inverted-U shaped.

This inverted-U pattern can be explained as follows: at low initial levels of competition (i.e high initial levels of $\delta(\chi)$), type-1 firms have strong reason to innovate; it follows that many intermediate sectors in the economy will end up being type-0 firms in steady-state (this we refer to as the *composition effect* of competition on the relative equilibrium fractions of type-0 and type-1); but then the dominant effect of competition on innovation is the escape competition effect whereby more competition fosters innovation by type-0 firms. On the other hand, at high initial levels of competition, innovation incentives in type-1 sectors are so low that a sector will remain of type-1 for a long time, and therefore many sectors will end up being of type-1 in steady-state, which in turn implies that the negative Schumpeterian appropriability effect of competition on innovation should tend to dominate in that case.

4.2 Empirical predictions

The above analysis generates several interesting predictions:

1. Innovation in sectors in which firms are close to the technology frontier, react positively to an increase in product market competition;
2. Innovation reacts less positively, or negatively, in sectors in which firms are further below the technological frontier;
3. The average fraction of frontier sectors decreases, i.e the average technological gap between incumbent firms and the frontier in their respective sectors increases, when competition increases;
4. The overall effect of competition on aggregate innovation, is inverted-U shaped.[23]

These predictions have been confronted by Aghion et al (2002) with UK firm level data on competition and patenting, and we briefly summarize their findings in the next subsection.

[23] Although perhaps only the second part of the inverse U will be observable. See footnote **??** above.

4.3 Empirical evidence and relationship to literature

Most innovation-based growth models -including the quality improvement model developed in the above two sections- would predict that product market competition is detrimental to growth as it reduces the monopoly rents that reward successful innovators (we refer to this as the Schumpeterian effect of competition). However, an increasing number of empirical studies have cast doubt on this prediction. The empirical IO literature on competition and innovation starts with the pioneering work of Scherer (1965), followed by Cohen-Levin (1967), and more recently by Geroski (1994). All these papers point to a positive correlation between competition and growth. However, competition is often measured by the inverse of market concentration, an indicator which Boone (2000) and others have shown to be problematic: namely, higher competition between firms with different unit costs may actually result in a higher equilibrium market share for the low cost firm! More recently, Nickell (1996) and Blundell et al (1999) have made further steps by conducting cross-industry analyses over longer time periods and by proposing several alternative measures of competition, in particular the inverse of the Lerner index (defined as the ratio of rents over value added) or by the number of competitors for each firm in the survey. However, none of these studies would uncover the reason(s) why competition can be growth-enhancing or why the Schumpeterian effect does not seem to operate.

It is by merging the Schumpeterian growth paradigm with previous patent race models (in which each of two incumbent firms would both, compete on the product market and innovate to acquire a lead over its competitor), that Aghion-Harris-Vickers (1997), henceforth AHV, and Aghion-Harris-Howitt-Vickers (2001), henceforth AHHV, have developed new models of competition and growth with step-by-step innovations that reconcile theory and evidence on the effects of competition and growth: by introducing the possibility that innovations be made by incumbent firms that compete "neck-and-neck", these extensions of the Schumpeterian growth framework show the existence of an "escape competition"effect that counteracts the Schumpeterian effect described above. What facilitated this merger between the Schumpeterian growth approach and the patent race models, is that: (i) both featured quality-improving innovations; (ii) models with vertical innovations in turn were particularly convenient to formalize the notion of technological distance and that of "neck-and-neck" competition. A main prediction of this new vintage of endogenous growth models, is that competition should be most growth-enhancing in sectors in which incumbent firms are close to the technological frontier and/or compete "neck-and-neck" with one another, as it is in those sectors that the "escape competition" effect should be the strongest.

These models in turn have provided a new pair of glasses for deeper empirical analyses of the relationship between competition/entry and innovation/growth. The two studies we briefly mention in the remaining part of this section have not only produce interesting new findings; they also suggested a whole new way of confronting endogenous growth theories with data, one that is more directly grounded on serious microeconometric analyses based on detailed firm/industry

panels.

The paper by Aghion-Bloom-Blundell-Griffith-Howitt (2002), henceforth ABBGH, takes a new look at the effects of product market competition on innovation, by confronting the main predictions of the AHV and AHHV models to firm level data. The prediction we want to emphasize here as it is very much in tune with our theoretical discussion in the previous subsections, is that the escape competition effect should be strongest in industries in which firms are closest to the technological frontier.

ABBGH considers a UK panel of individual companies during the period 1968-1997. This panel includes all companies quoted on the London Stock Exchange over that period, and whose names begin with a letter from A to L. To compute competition measures, the study uses firm level accounting data from Datastream; product market competition is in turn measured by one minus the Lerner index (ratio of operating profits minus financial costs over sales), controlling for capital depreciation, advertising expenditures, and firm size. Furthermore, to control for the possibility that variations in the Lerner index be mostly due to variations in fixed costs, we use policy instruments such as the implementation of the Single Market Program or lagged values of the Lerner index as instrumental variables. Innovation activities, in turn, are measured both, by the number of patents weighted by citations, and by R&D spending. Patenting information comes from the US Patent Office where most firms that engage in international trade register their patents; in particular, this includes 461 companies on the London Stock Exchange with names starting by A to L, for which we already had detailed accounting data. Finally, technological frontier is measured as follows: suppose a UK firm (call it i) belongs to some industry A; then we measure technological distance by the difference between the maximum TFP in industry A across all OECD countries (we call it TFP_F , where the subscript "F" refers to the technological frontier) and the TFP of the UK firm, divided by the former:

$$m_i = \frac{TFP_F - TFP_i}{TFP_F}.$$

Figure 1 summarizes our main findings.

FIGURE 1 *HERE*

Each point on this figure corresponds to one firm in a given year. The upper curve considers only those firms in industries where the average distance to the technological frontier is less than the median distance across all industries, whereas the lower curve includes firms in all industries. We clearly see that the effect of product market competition on innovation is all the more positive that firms are closer to the technological frontier (or equivalently are more "neck-and-neck"). Another interesting finding that comes out of the Figure, is that the Schumpeterian effect is also at work, and that it dominates at high initial levels of product market competition. This in turn reflects the "composition effect" pointed out in the previous subsection: namely, as competition increases and

neck-and-neck firms therefore engage in more intense innovation to escape competition, the equilibrium fraction of neck-and-neck industries tends to decrease (equivalently, any individual firm spends less time in neck-and-neck competition with its main rivals) and therefore the average impact of the escape competition effect decreases at the expense of the counteracting Schumpeterian effect. The ABBGH paper indeed shows that the average distance to the technological distance increases with the degree of product market competition. The Schumpeterian effect was missed by previous empirical studies, mainly as a result of their being confined to linear estimations. Instead, more in line with the Poisson technology that governs the arrival of innovations both, in Schumpeterian and in patent race models, ABBGH use a semi-parametric estimation method in which the expected flow of innovations is a piecewise polynomial function of the Lerner index.

4.4 A remark on inequality and growth

Our discussion of the effects of competition on growth also sheds light on the current debate on the effects of income or wealth inequality on growth. A recent literature[24] has emphasized the idea that in an economy with credit-constraints, where the poor do not have full access to efficient investment opportunities; redistribution may enhance investment by the poor more than it reduces incentives for the rich, thereby resulting in higher aggregate productive efficiency in steady-state and higher rate of capital accumulation on the transition path to the steady-state. Our discussion of the effects of competition on innovation and growth, hints at yet another negative effect of excessive wealth concentration on growth: to the extent that innovative activities tend to be more intense in sectors in which firms or individuals compete "neck-and-neck", taxing further capital gains by firms that are already well ahead of their rivals in the same sector, may enhance the aggregate rate of innovation by shifting the overall distribution of technological gaps in the economy towards a higher fraction of neck-and-neck sectors in steady-state.

More generally, having too many sectors in which technological knowledge and/or wealth are highly concentrated, may inhibit growth as it both, discourages laggard firms or potential entrants, and reduces the leader's incentives to innovate in order to escape competition given that the competitive threat coming from laggards or potential entrants is weak; the leader may actually prefer to invest her wealth into entry deterrence activities. These considerations may in turn explain why, following a high growth period during the industrial revolution in the 19th century, growth slowed down at the turn of the 20th century in France or England at the same time wealth distribution became highly concentrated: the high concentration of wealth that resulted from the industrial revolution, turned the innovators of the mid 19th century into entrenched incumbents with the power to protect their dominant position against competition by new potential entrants.[25]

[24] For example, see Galor-Zeira (1993), Banerjee-Newman (1993), and Aghion-Bolton (1997).
[25] See Piketty et al (2003).

5 Scale effects[26]

5.1 Theory

Jones (1995) has pointed out that the simple model of the preceding sections whereby increased population leads to increased growth, by raising the size of the market for a successful entrepreneur and by raising the number of potential R&D workers, is not consistent with post-war evidence. In the United States, for example, the number of scientists and engineers engaged in R&D has grown by a factor of five since the 1950s with no significant trend increase in productivity growth. This refutes the version of the basic model in which productivity growth is a function of skilled labor applied to R&D (section 2.3 above). Likewise, the fact that productivity-adjusted R&D has grown substantially over the same period rejects the version of the model presented in section 2 above in which productivity growth is a function of productivity-adjusted research.

5.1.1 The Schumpeterian (fully endogenous) solution

Schumpeterian theory deals with this problem of the missing scale effect on productivity growth by incorporating Young's (1998) insight that as an economy grows, proliferation of product varieties reduces the effectiveness of R&D aimed at quality improvement, by causing it to be spread more thinly over a larger number of different sectors.[27] When modified this way the theory is consistent with the observed coexistence of stationary TFP growth and rising R&D input, because in a steady state the growth-enhancing effect of rising R&D input is just offset by the deleterious effect of product proliferation.

The simplest way to illustrate this modification is to suppose that the number of sectors m is proportional to the size of population L. For simplicity normalize so that $m = L$.[28] Then the growth equation (10) becomes:

$$g = \lambda^2 \delta(\chi)(\gamma - 1) \tag{23}$$

It follows directly from comparing (23) with (10) that all the comparative-statics propositions of section 2.4 above continue to hold except that now the growth rate is independent of population size.

5.1.2 The semi-endogenous solution

Jones (1999) argues that this resolution of the problem is less intuitively appealing than his alternative semi-endogenous theory, built on the idea of diminishing returns to the stock of knowledge in R&D. In this theory sustained *growth* in R&D input is necessary just to maintain a given rate of productivity growth.

[26] This section drawson Ha and Howitt (2004).

[27] Variants of this idea have been explored by van de Klundert and Smulders (1997), Peretto (1998), Dinopoulos and Thompson (1998) and Howitt (1999).

[28] Thus, in contrast to Romer (1990) where horizontal innovations drive the growth process, here product proliferation eliminates scale effects whereas long-run growth is still ultimately driven by quality-improving innovations.

Semi-endogenous growth theory has a stark long-run prediction, namely that the long-run rate of productivity growth, and hence the long-run growth rate of per-capita income, depend on the rate of population growth, which ultimately limits the growth rate of R&D labor, to the exclusion of all economic determinants.

In Jones's formulation:

$$g = \lambda f(n) A^{\phi-1} (\gamma - 1), \qquad \phi < 1$$

where the R&D input n is measured by the number R&D workers in G5 countries. Except for the assumption of diminishing returns ($\phi < 1$) this is equivalent to the original formulation (5) above. In the special case where f takes a Cobb-Douglas form we have, in continuous time:

$$g \equiv \dot{A}/A = \lambda n^{\sigma} A^{\phi-1} (\gamma - 1)$$

so that:

$$\dot{g}/g = (1 - \phi)(\gamma' g_n - g) \tag{24}$$

where $g_n = \dot{n}/n$ is the growth rate of R&D workers and $\gamma' = \sigma/(1 - \phi)$.

This semi-endogenous model is compatible with the observation of positive trend growth in R&D input, because as long as $\phi < 1$ and the time path of g_n is bounded, the differential equation (24) yields a bounded solution for productivity growth. In particular, if g_n is constant, or approaches a constant, then

$$g \to \gamma' g_n$$

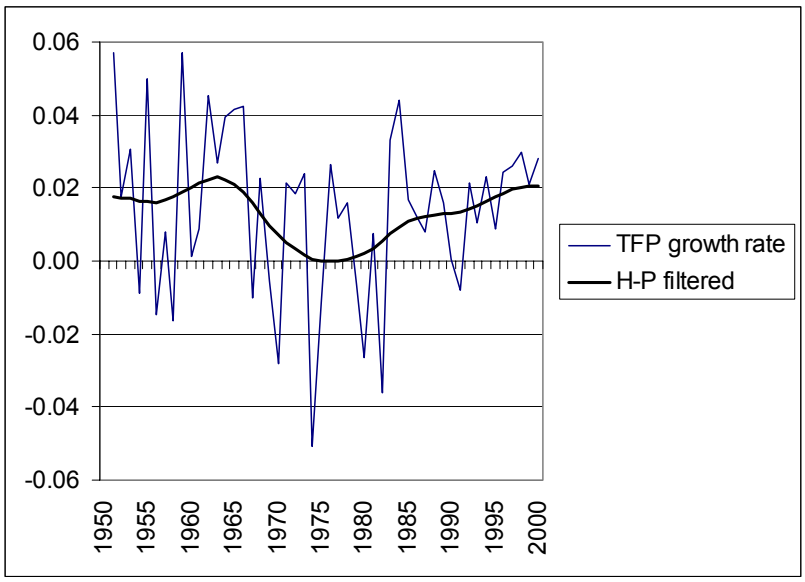
In the long run the growth rate of R&D labor cannot exceed the growth rate η of population, and in a balanced growth equilibrium it will equal η. Likewise, the growth rate of productivity-adjusted R&D expenditure will equal η along a balanced growth path. Hence the radical implication that the long-run growth rate of an economy will equal $\gamma'\eta$, independently of what fraction of society's resources are is assigned to knowledge creation. Policies to stimulate R&D will have at most transitory effects on productivity growth and, by extension, on per-capita income growth.

5.2 Evidence

These two competing approaches to reconciling R&D-based theory with the observed upward trend in R&D input offer a stark contrast. The Schumpeterian approach with product-proliferation effects retains all the characteristic comparative statics predictions of endogenous growth theory as outlined in section 2.4 above, while Jones's semi-endogenous theory denies all these predictions.

Fortunately the two competing approaches can also be tested using observed trends in productivity growth and R&D input. Specifically, the semi-endogenous model implies that the growth rate of productivity will track the growth rate

Figure 2: TFP growth rates, US, 1950-2000



of R&D input, whereas the Schumpeterian model implies that it will track the fraction of GDP spent on R&D.[29]

To derive this Schumpeterian implication note that, according to the growth equation (5), productivity growth depends on productivity-adjusted R&D per sector, n. Given the assumption $m = L$, if GDP per person grows asymptotically at the rate g then n will be proportional to the fraction of GDP spent on R&D.

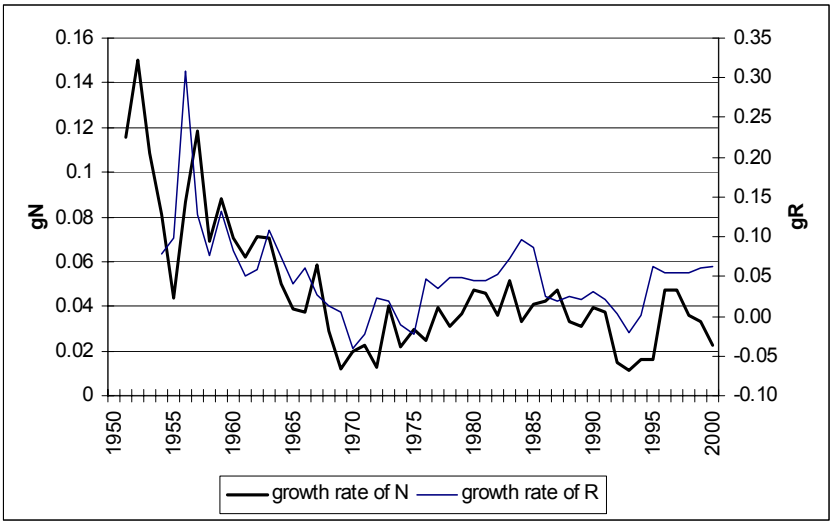
Figure 2 shows the growth rate of productivity in the United States from 1950 to 2000. There is no discernible trend. An Augmented Dickey-Fuller test rejects a unit root at the 1% significance level, confirming the stationarity of this series. Thus semi-endogenous theory implies that the growth rate of R&D input should also be trendless and stationary, whereas Schumpeterian theory implies that the R&D/GDP ratio should be trendless and stationary.

5.2.1 Results

Figure 3 shows that growth rates of the number of R&D workers in the G5 countries, N, and US R&D expenditure, R, appear to have a substantial negative trend, having fallen roughly fourfold since the early 1950s. The impression of

[29] Zachariadis (2003) shows that the fully-endogenous Schumpeterian theory without scale effects also passes a number of other tests using U.S. data. Specifically, he finds using two-digit industry level data that patenting, technological progress and productivity growth all depend upon the ratio of R&D expenditures to output, as implied by the fully endogenous theory.

Figure 3: Trend of growth rates for G5 R&D workers and US R&D expenditures



non-stationarity is supported by an Augmented Dickey-Fuller test, which fails to reject a unit root in g_N at the 5% level.

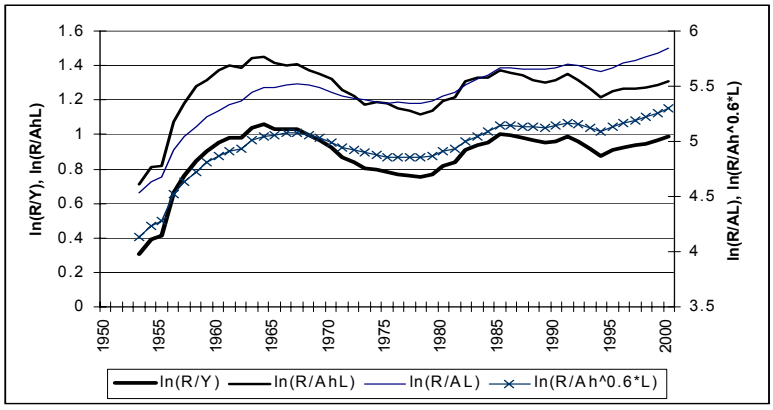
These findings are inconsistent with the implications of semi-endogenous growth theory.[30] Indeed they undermine the central proposition of semi-endogenous theory, because if productivity growth can be sustained for 50 years in the face of such a large fall in the growth rate of R&D labor then there is no reason to suppose that population growth limits productivity growth, except perhaps over a time scale of hundreds of years.

Figure 4 shows that the fraction of GDP spend on R&D in the US looks more or less stable with perhaps a small upward trend.[31] It is notable that ever since 1957, R&D as a percentage of GDP has been fluctuating between 2.1% and 2.9%, with similar movements as in productivity growth: downward trend for 1964-1975 and upward trend for 1975-2000. The stationarity of this

[30] The data on G5 R&D workers come from Jones, who had to guess at the non-US component from 1950 to 1965. However, Ha and Howitt (2004) consider a broader range of R&D measures. They also show that the formal cointegration predictions implied by semi-endogenous theory are not found in these data, even if attention is restricted to the post-1965 date, while the even tighter cointegration predictions implied by Schumpeterian theory are found. Ha and Howitt also conduct a calibration exercise and show that the semi-endogenous model fits the data of US productivity growth best when ϕ is almost equal to unity as in the fully endogenous model.

[31] There appears to be a more significant upward trend if we omit space and defense R&D, as is done by many researchers in the productivity literature on the grounds that they do not find spillovers from these components of R&D. However, this literature has not allowed for the very long lags with which we think federal R&D has its effects. Moreover, throwing out federal R&D would at times amount to throwing out about 70% of the total.

Figure 4: Trend of R&D intensity, US (log)



series is confirmed by an Augmented Dickey-Fuller test, which rejects a unit root at the 1% level. This is in conformity with the version of Schumpeterian theory presented above, adjusted to take into account the effects of product proliferation.

5.3 Concluding remarks

The scale effect whereby increased population should lead to increased productivity growth clearly refutes a simple interpretation of the model in section 2 above, in which L stands for the *number* of (skilled) individuals. However, we have shown in this section that even if we stick to this interpretation of L, a simple variant of the Schumpeterian model can be developed, which carries all the same long-run growth implications except for the scale effect. The rival semi-endogenous theory of Jones (1995), which denies endogenous growth in the very long run, is inconsistent with the observation that productivity growth can be sustained through half a century of falling growth in R&D labor. The analogous implication of amended Schumpeterian theory, namely that productivity growth can be sustained as long as society allocates a constant fraction of its resources to research, is consistent with the evidence.

Two brief remarks conclude this section. First, there is no evidence pointing to the absence of a scale effect at the world level or in small closed economies. That the stock of educated labor should affect technological convergence and productivity growth worldwide was first pointed out by Nelson and Phelps (1966). Second, if we replace L by Le^N, where e^N denotes the *quality* of the labor force as measured for example by the average number of years in schooling (so that more educated countries have more efficiency units of labor), then even

if one eliminates scale effects by taking $L = m$, there will still remain a "level effect" embodied in the e^N term, whereby a higher average number of years of education N has a positive effect on growth. In the next section we show that increasing the fraction of highly educated workers and/or increasing the average number of years in schooling, have a positive impact on the rate of productivity growth, but the extent of which depends upon the country's distance to the world technology frontier: in particular, the closer a country is to the frontier, the higher is the effect of an additional year of higher education on its rate of productivity growth.

6 Linking growth to institutional change

6.1 From Schumpeter to Gerschenkron

By linking growth to innovation and entrepreneurship, and innovation incentives in turn to characteristics of the economic environment, new growth theories made it possible to analyze the interplay between growth and the design of policies and institutions. For example, the basic model developed in Section 2 suggested that long-run growth would be best enhanced by a combination of good property right protection (to protect the rents of innovators against imitation), a good education system (to increase the efficiency of R&D activities and/or the supply of skilled manufacturing labor), and a stable macroeconomy to reduce interest rates (and thereby increase the net present value of innovative rents). Our discussion of convergence clubs in Section 3 then suggested that the same policies or institutions would also increase a country's ability to join the convergence club.

Now, new growth theories may be criticized by development economists and policy makers, precisely because of the universal nature of the policy recommendations that appear to follow from them: no matter how developed a country or sector currently is, it seems that one should prescribe the same medicines (legal reform to enforce property rights, investment climate favorable to entrepreneurship, education, macrostability,..) to maximize the growth prospects of that country or sector.

However, in his essay on *Economic Backwardness in Historical Perspective,* Gerschenkron (1962) argues that relatively backward economies could more rapidly catch up with more advanced countries by introducing "appropriate institutions" that are growth-enhancing at an early stage of development but may cease to be so at a later stage. Thus, countries like Japan or Korea managed to achieve very high growth rates between 1945 up until the 1990s with institutional arrangements involving long-term relationships between firms and banks, the predominance of large conglomerates, and strong government intervention through export promotion and subsidized loans to the enterprise sector, all of which depart from the more market-based and laissez-faire institutional model pioneered and promoted by the US.

That growth-enhancing institutions or policies might change with a coun-

try's or sector's distance to the technological frontier, should not come as a total surprise to our readers at this point: in the previous section, we saw that competition could have opposite effects on innovation incentives depending on whether firms were initially closer to or farther below the fringe in the corresponding industry (it would enhance innovations in neck-and-neck industries, and discourage it in industries where innovating firms are far below the frontier). The same type of conclusion turns out to hold true when one looks at the interplay between countries' distance to the world technology frontier and "openness". Using a cross-country panel of more than 100 countries over the 1960-2000 period, Acemoglu-Aghion-Zilibotti (2002), henceforth AAZ, regress the average growth rate over a five year period on a country's distance to the US frontier (measured by the ratio of GDP per capita in that country to per capita GDP in the US) at the beginning of the period. Then, splitting the sample of countries in two groups, corresponding respectively to a high and a low openness group according to Frankel-Romer's openness indicator, AAZ show that average growth decreases more rapidly as a country approaches the world frontier when openness is low. Thus, while a low degree of openness does not appear to be detrimental to growth in countries far below the world frontier, it becomes increasingly detrimental to growth as the country approaches the frontier. AAZ repeat the same exercise using entry costs to new firms (measured as in Djankov et al (2001)) instead of openness, and they obtain a similar conclusion, namely that high entry costs are most damaging to growth when a country is close to the world frontier, unlike in countries far below the frontier.

In this section, we shall argue that Gerschenkron's idea of "appropriate institutions" can be easily embedded into our growth framework, in a way that can help substantiate the following claims:

1. different institution or policy design affects productivity growth differently depending upon a country's or sector's distance to the technological frontier;

2. a country's distance to the technological frontier affect the type of organizations we observe in this country (e.g, bank versus market finance, vertical integration versus outsourcing,..).

The remaining part of the section is organized as follows. We first describe the growth equation which AAZ introduce to embed the notion of "appropriate institutions" into the above growth framework. We then focus on the first question about the effects of institution design on productivity growth, by concentrating on the relationship between growth and the organization of education. Finally we briefly discuss the effects of distance on equilibrium institutions in a concluding subsection.

6.2 A simple model of appropriate institutions

Consider the following variant of the multi-country growth model of Section 3. In each country, a unique general good which also serves as numéraire, is

produced competitively using a continuum of intermediate inputs according to:

$$y_t = \int_0^1 \left(A_t\left(i\right)\right)^{1-\alpha} x_t\left(i\right)^{\alpha} \; di, \tag{25}$$

where $A_t\left(i\right)$ is the productivity in sector i at time t, $x_t(i)$ is the flow of intermediate good i used in general good production again at time t, and $\alpha \in [0,1]$.

As before, ex post each intermediate good producer faces a competitive fringe of imitators that forces her to charge a limit price $p_t\left(i\right) = \chi > 1$. Consequently, equilibrium monopoly profits (gross of the fixed cost) are simply given by::

$$\pi_t\left(i\right) = \delta A_t\left(i\right)$$

where $\delta \equiv \left(\chi - 1\right)\chi^{-\frac{1}{1-\alpha}}$.

We still let

$$A_t \equiv \int_0^1 A_t\left(i\right) di$$

denote the average productivity in the country at date t, $\overline{A}_t$ the productivity at the world frontier which we assume to grow at the constant rate g from one period to the next, and $a_t = A_t/\bar{A}_t$ the (inverse) measure of the country's distance to the technological frontier at date t.

The main departure from the convergence model in Section 3, lies in the equation for productivity growth. Suppose that intermediate firms have two ways to generate productivity growth: (a) they can imitate existing world frontier technologies; (b) they can innovate upon the previous local technology. More specifically, we assume:

$$A_t\left(i\right) = \eta \bar{A}_{t-1} + \gamma A_{t-1}, \tag{26}$$

where $\eta \bar{A}_{t-1}$ and γA_{t-1} refer respectively to the imitation and innovation components of productivity growth. Imitations use the existing frontier technology at the end of period $(t-1)$, thus they multiply $\bar{A}_{t-1}$, whereas innovations build on the knowledge stock of the country, and therefore they multiply A_{t-1}.

Now dividing both sides of (26) by $\overline{A}_t$, using the fact that

$$\overline{A}_t = (1+g)\bar{A}_{t-1},$$

and integrating over all intermediate sectors i, we immediately obtain the following linear relationship between the country's distance to frontier a_t at date t and the distance to frontier a_{t-1} at date $t-1$:

$$a_t = \frac{1}{1+g}(\eta + \gamma a_{t-1}). \tag{27}$$

This equation clearly shows that the relative importance of innovation for productivity growth, increases as: (i) the country moves closer to the world technological frontier, i.e as a_{t-1} moves closer to 1, whereas imitation is more important when the country is far below the frontier, i.e when a_{t-1} is close to

zero; (ii) a new technological revolution (e.g the ITC revolution) occurs that increases the importance of innovation, i.e increases γ.

This immediately generates a theory of "appropriate institutions" and growth: suppose that imitation and innovation activities do not require the same institutions. Typically, imitation activities (i.e η in the above equation (27)) will be enhanced by long-term investments within (large) existing firms, which in turn may benefit from long-term bank finance and/or subsidized credit as in Japan or Korea since 1945. On the other hand, innovation activities (i.e γ) require initiative, risk-taking, and also the selection of good projects and talents and the weeding out of projects that turn out not to be profitable. This in turn calls for more market-based and flexible institutions, in particular for a higher reliance on market finance and speculative monitoring, higher competition and trade liberalization to weed out the bad projects, more flexible labor markets for firms to select the most talented or best matched employees, non-integrated firms to increase initiative and entrepreneurship downstream, etc. It then follows from equation (27) that the growth-maximizing institutions will evolve as a country moves towards the world technological frontier. Far below the frontier, a country will grow faster if it adopts what AAZ refers to as *investment-based* institutions or policies, whereas closer to the frontier growth will be maximized if the country switches to *innovation-based* institutions or policies.

A natural question is of course whether institutions actually change when they should from a growth- (or welfare-) maximizing point of view, in other words how do equilibrium institutions at all stages of development compare with the growth-maximizing institutions? This question is addressed in details in AAZ, and we will come back to it briefly in the last subsection.

6.3 Appropriate education systems

In his seminal paper on economic development, Lucas (1988) emphasized the *accumulation* of human capital as a main engine of growth; thus, according to the analysis in that paper, cross-country differences in growth rates across countries would be primarily attributable to differences in *rates of accumulation* of human capital. An alternative approach, pioneered by Nelson-Phelps (1966), revived by the Schumpeterian growth literature[32], would instead emphasize the combined effect of the *stock* of human capital and of the innovation process in generating long-run growth and fostering convergence. In this alternative approach, differences in growth rates across countries would be mainly attributable to differences in *stocks* of human capital, as those condition countries' ability to innovate or to adapt to new technologies and thereby catch up with the world technological frontier. Thus, in the basic model of Section 2, the equilibrium R&D investment and therefore the steady-state growth rate were shown to be increasing in the aggregate supply of (skilled) labor L and in the productivity of research λ, both of which refer more to the *stock* and *efficiency* of human capital than to its rate of accumulation.

[32] For example, see Acemoglu (1996, 2002), Aghion-Howitt-Violante (2002) and Aghion (2002).

Now, whichever approach one takes, and the evidence so far supports the two approaches as being somewhat complementary, once again one may worry about growth models delivering too general a message, namely that more education is always growth enhancing. In this subsection we will try to go one step further and argue that the AAZ specification (summarized by the above equation (26), can be used to analyze the effects, not only of the total *amount* of education, but more importantly of the *organization* of education, on growth in countries at different stages of development..

This subsection, which is based on Vandenbussche-Aghion-Meghir (2003), henceforth VAM, focuses on one particular aspect of the organization of education systems, namely the mix between primary, secondary, and higher education. We consider a variant of the AAZ model outlined in the previous subsection, in which innovation requires highly educated labor, whereas imitation can be performed by both, highly educated and lower-skill workers. A main prediction emerging from this a model, is that the closer a country gets to the world technology frontier, the more growth-enhancing it becomes to invest in higher education. In the latter part of the subsection we confront this prediction with preliminary cross-country evidence.

6.3.1 Distance to frontier and the growth impact of higher education

There is again a unique general good, produced competitively using a continuum of intermediate inputs according to:

$$y = \int_0^1 A(i)^{1-\alpha} x(i)^{\alpha} di, \tag{28}$$

where $A(i)$ is the productivity in sector i, $x(i)$ is the flow of intermediate good i used in general good production, $\alpha \in [0,1]$.

In each intermediate sector i, one intermediate producer can produce the intermediate good with leading-edge productivity $A_t(i)$, using general good as capital according to a one-for-one technology. As before, ex post each intermediate good producer faces a competitive fringe of imitators that forces her to charge a limit price $p(i) = \chi > 1$. Consequently, we have:

$$p(i) = \chi = \frac{\partial y}{\partial x},$$

so that equilibrium monopoly profits in each sector i are given by::

$$\pi(i) = (p(i) - 1)x(i) = \delta\pi(i) = \delta A(i)L$$

where $\delta = (\chi - 1)(\frac{\chi}{\alpha})^{\frac{-1}{1-\alpha}}$.

As in the previous subsection, intermediate firms can increase productivity, either by imitating frontier technologies or by innovating upon existing technologies in the country. Imitation can be performed by both types of workers,

whereas innovation requires high education. More specifically, we focus on the following class of productivity growth functions:

$$A_t(i) - A_{t-1}(i) = u_{m,i,t}^{\sigma} s_{m,i,t}^{1-\sigma} \bar{A}_{t-1} + \gamma u_{n,i,t}^{\phi} s_{n,i,t}^{1-\phi} A_{t-1}, \tag{29}$$

where $u_{m,i,t}$ (resp. $s_{m,i,t}$) is the amount of unskilled (resp. skilled) labor used in imitation in sector i at time t, $u_{n,i,t}$ (resp. $s_{n,i,t}$) is the amount of unskilled (resp. skilled) units of labor used by sector i in innovation at time t, σ (resp. ϕ) is the elasticity of unskilled labor in imitation (resp. innovation), and $\gamma > 0$ measures the relative efficiency of innovation compared to imitation in generating productivity growth.

We shall assume:

(A1) *The elasticity of skilled labor is higher in innovation than in imitation, and conversely for the elasticity of unskilled labor, that is:* $\phi < \sigma$.

Let S (resp. $U = 1 - S$) denote the fraction of the labor force with higher (resp. primary or secondary) education. Let $w_{u,t}\bar{A}_{t-1}$ (resp. $w_{s,t}\bar{A}_{t-1}$) denote the current price of unskilled (resp. skilled) labor.

The total labor cost of productivity improvement by intermediate firm i at time t, is equal to:

$$W_{i,t} = [w_{u,t}(u_{m,i,t} + u_{n,i,t}) + w_{s,t}(s_{m,i,t} + s_{n,i,t})]\overline{A}_{t-1}.$$

Letting $a_t = A_t/\bar{A}_t$ measure the country's distance to the technological frontier, and letting the frontier technology $\bar{A}_t$ grow at constant rate g, the intermediate producer will solve:

$$\max_{u_{m,i,t}, s_{m,i,t}, u_{n,i,t}, s_{n,i,t}} \{\delta[u_{m,i,t}^{\sigma} s_{m,i,t}^{1-\sigma}(1 - a_{t-1}) + \gamma u_{n,i,t}^{\phi} s_{n,i,t}^{1-\phi} a_{t-1}]\bar{A}_{t-1} - W_{i,t}\}. \tag{30}$$

Using the fact that all intermediate firms face the same maximization problem, and that there is a unit mass of intermediate firms, we necessarily have:

$$u_{j,i,t} \equiv u_{j,t}; s_{j,i,t} \equiv s_{j,t} \text{ for all } i \text{ and for } j = m, n; \tag{31}$$

and

$$S = s_{m,t} + s_{n,t}; U = 1 - S = u_{m,t} + u_{n,t}. \tag{32}$$

Taking first order conditions for the maximization problem (30), then making use of (31) and (32), and then computing the equilibrium rate of productivity growth

$$g_t = \int_0^1 \frac{A_t(i) - A_{t-1}}{A_{t-1}} di,$$

one can establish (see VAM (2003)):

Lemma 1 *Let* $\psi = \frac{\sigma(1-\phi)}{(1-\sigma)\phi}$. *If parameter values are such that the solution to (30) is interior, then we have:*

$$\frac{\partial g_t}{\partial a} = \phi(1-\phi)h'(a)h(a)^{1-\phi}[h(a)U - S],$$

where

$$h(a) = (\frac{(1-\sigma)\psi^{\sigma}(1-a)}{(1-\phi)\gamma a})^{\frac{1}{\sigma-\phi}} \geq \frac{S}{U}.$$

This, together with the fact that $h(a)$ is obviously decreasing in a given our assumption (A1), immediately implies:

Proposition 2 *A marginal increase in the fraction of labor with higher education, enhances productivity growth all the more the closer the country is from the world technology frontier, that is:*

$$\frac{\partial^2 g_t}{\partial a \partial S} > 0.$$

The intuition follows directly from the Rybczynski theorem in international trade. Stated in the context of a two sector-two input economy, this theorem says that an increase in the supply of input in the sector that uses that input more intensively, should increase "output" in that sector more than proportionally. To transpose this result to the context of our model, consider the effect of an increase in the supply of skilled labor, keeping the supply of unskilled labor fixed and for given a. Given that skilled workers contribute relatively more to productivity growth and profits if employed in innovation rather than in imitation (our Assumption (A1)), the demand for additional skilled labor will tend to be higher in innovation. But then the marginal productivity of unskilled labor should also increase more in innovation than in imitation, hence a net flow of unskilled workers should also move from imitation into innovation. This in turn will enhance further the marginal productivity of skilled labor in innovation, thereby inducing an ever greater fraction of skilled labor to move to innovation. Now the closer the country is to the technology frontier (i.e the higher a), the stronger this Rybszynski effect as a higher a increases the efficiency of both, skilled and unskilled labor, in innovation relative to imitation. A second, reinforcing, reason is that an increase in the fraction of skilled labor reduces the amount of unskilled labor available in the economy, hence reducing the marginal productivity of skilled labor in imitation, all the more the closer the country is from the frontier.

We can now confront this prediction with cross-country evidence on higher education, distance to frontier, and productivity growth.

6.3.2 Empirical evidence

The prediction that higher education has stronger growth-enhancing effects close to the technological frontier can be tested using cross-regional or cross-country data. Thus VAM consider a panel dataset of 19 OECD countries over the period 1960-2000. Output and investment data are drawn from Penn World Tables 6.1 (2002) and human capital data from Barro-Lee (2000). The Barro-Lee data indicate the fraction of a country's population that has reached a certain level of schooling at intervals of five years, so VAM use the fraction that has received

some higher education together with their measure of TFP (itself constructed assuming a constant labor share of .7 across countries) to perform the following regression:

$$g_{j,t} = \alpha_{0,j} + \alpha_1 dist_{j,t-1} + \alpha_2 \Lambda_{j,t-1} + \alpha_3 (dist_{j,t-1} * \Lambda_{j,t-1}) + u_{j,t},$$

where $g_{j,t}$ is country j's growth rate over a five year period, $dist_{j,t-1}$ is country j's closeness to the technological frontier at $t-1$ (i.e. 5 years before), $\Lambda_{j,t-1}$ is the fraction of the working age population with some higher education in the previous period and $\alpha_{0,j}$ is a country dummy controlling for country fixed effects. The closeness variable is instrumented with its lagged value at $t-2$, and the fraction variable is instrumented using expenditure on tertiary education per capita lagged by two periods, and the interaction term is instrumented using the interaction between the two instruments for closeness and for the fraction variables. Finally, the standard errors we report allow for serial correlation and heteroskedasticity.

The results from this regression are shown in Table 1 below. In particular, we find a positive and significant interaction between our education measure and closeness to the frontier, as predicted by the theory in the previous subsection. This result demonstrates that it is more important to expand years of higher education close to the technological frontier.

7 Conclusion

In this chapter we argued that the endogenous growth model with quality-improving innovations provides a framework for analyzing the determinants of long-run growth and convergence that is versatile, simple and empirically useful. Versatile, as the same framework can be used to analyze how growth interacts with development and cross-country convergence and divergence, how it interacts with industrial organization and in particular market structure, and how it interacts with organizations and institutional change. Simple, since all these aspects can be analyzed using the same elementary model. Empirically useful, as the framework generates a whole range of new microeconomic and macroeconomic predictions while it addresses empirical criticisms raised by other endogenous growth models in the literature.

Far from closing the field, the chapter suggests many avenues for future research. For example, on growth and convergence, more research remains to be done to identify the main determinants of cross-country convergence and divergence.[33] Also important, is to analyze the role of international intellectual property right protections and foreign direct investment in preventing or favoring convergence. On growth and industrial organization, we have restricted

[33] In Aghion, Howitt and Mayer-Foulkes (2004) we emphasize the role of credit constraints in R&D as a distinguishing factor between the countries that converge in growth rates and in levels towards the frontier, those that converge only in growth rates, and those that follow a divergent path towards a lower rate of long-run growth. Whether credit constraints, or other factors such as health, education, and property rights protection, are key to this three-fold classification, remains an open question

attention to product market competition among existing firms. But what can we say about entry and its impact on incumbents' innovation activities?[34] On institutions, we have just touched upon the question of how technical change interacts with organizational change. Do countries or firms/sectors actually get stuck in institutional traps of the kind described in Section 6? What enables such traps to disappear over time? How do political economy considerations interact with this process? There is also the whole issue of wage inequality and its interplay with technical change, on which the Schumpeterian approach developed in this chapter can also shed light.[35]

If we just had to select three aspects or questions, so far largely open, and which could also be explored using our approach, we would suggest the following. First, on the role of basic science in generating (very) long-term growth. Do fundamental innovations (or the so called "general purpose technologies") require the same incentive system and the same rewards as industrial innovations? How can one design incentive systems in universities so that university research would best complement private research? A second aspect is the interplay between growth and volatility. Is R&D and innovation procyclical or countercyclical, and is macroeconomic volatility always detrimental to innovation and growth? Answering this question in turn opens up a whole new research topic on the macropolicy of growth[36] A third aspect is the extent to which our growth paradigm can be applied to less developed economies. In particular, can we use the new growth approach developed in this chapter to revisit the important issue of poverty reduction?[37] On all these questions, we believe that over time compelling answers will emerge from a fruitful dialogue between applied theorists, in particular those working on endogenous growth models of the kind developed in this chapter, and microeconometricians who use firm-level panel data to analyze the interplay between competition and innovation or between productivity growth and organizations.

Finally, in this chapter we have argued that modelling growth as resulting from quality-improving innovations, provides a natural framework to address a whole array of issues from competition to development, each time with theoretical predictions that can be empirically tested and also lead to more precise policy prescriptions. However, one might think of more direct ways of testing the quality-ladder model against the variety model analyzed in the other chapters. For example, in current work with Pol Antras and Susanne Prantl, we are using a panel data set of UK firms over the past fifteen years, to assess whether variety had any impact on innovation and growth. Using input-output tables, our preliminary results suggest that exit of input firms has but a positive effect on the productivity growth of final producers.

[34] See Aghion et al (2003a, 2003b) for preliminary work on entry and growth.

[35] E.g, see Aghion (2003) and the chapter by Krusell and Violante in this Handbook volume.

[36] See Aghion-Angeletos-Banerjee-Manova (2004).

[37] See Aghion and Armendariz de Aghion (2004) for some preliminary thoughts on this aspect.

References

Acemoglu, D, Aghion, P, Griffith, R, and F. Zilibotti (2003): "Technology, Hold-Up, and Vertical Integration: What Do We Learn From Micro Data?", mimeo IFS, London.

Acemoglu, D, Aghion, P, and F. Zilibotti (2002): "Distance to Frontier, Selection, and Economic Growth", NBER Working Paper 9066

Aghion, P, Angeletos, M, Banerjee, A, and K. Manova (2004), " Volatility and Growth: Financial Development and the Cyclical Composition of Investment", Harvard Working Paper.

Aghion, P, and B. Armendariz de Aghion (2004): "A New Growth Approach to Poverty Reduction", mimeo Harvard.

Aghion, P, A. V. Banerjee, and T. Piketty (1999): "Dualism and Macroeconomic Volatility", *Quarterly Journal of Economics,* 114 pp. 1359-97.

Aghion, P, Bloom, B, Blundell, R, Griffith, R, and P. Howitt (2002): "Competition and Innovation: An Inverted-U Relationship", NBER Working Paper 9269.

Aghion, P, and P. Bolton (1997): "A Model of Trickle-Down Growth and Development", *Review of Economic Studies.*

Aghion, P, Harris, C, Howitt, P, and J. Vickers (2001): "Competition, Imitation, and Growth with Step-by-Step Innovation", *Review of Economic Studies,* 68, 467-492.

Aghion, P, Harris, C, and J. Vickers (1997): "Competition and Growth with Step-by-Step Innovation: An Example", *European Economic Review Papers and Proceedings,* pp. 771-782

Aghion, P, and P. Howitt (1992): "A Model of Growth through Creative Destruction", *Econometrica,* 60, 323-351.

Aghion, P, and P. Howitt (1998): *Endogenous Growth Theory,* MIT Press, Cambridge, MA.

Aghion, P, Howitt, P, and D. Mayer-Foulkes (2004): "The Effect of Financial Development on Convergence: Theory and Evidence", unpublished, Brown University.

Aghion, P, P. Howitt, and G. L. Violante (2002): "General Purpose Technology and Wage Inequality", *Journal of Economic Growth,* 7, pp. 315-45.

Aghion, P, Meghir, C, and J. Vandenbussche (2003): "Productivity Growth and the Composition of Education Spending", unpublished.

Aghion, P, and J. Tirole (1997): "Formal and Real Authority in Organizations", *Journal of Political Economy,* 105, 1-29.

Arrow, K. J. (1969): "Classificatory Notes on the Production and Transmission of Technological Knowledge", *American Economic Review Papers and Proceedings,* 59, pp. 29-35.

Banerjee, A, and A. Newman (1993): "Occupational Choice and the Process of Development", *Journal of Political Economy,* 101, 274-298.

Barro, R. J., and X. Sala-i-Martin (1992): "Convergence", *Journal of Political Economy,* 100, pp. 223-51.

Barro, R. J., and X. Sala-i-Martin (1997): "Technological Diffusion, Convergence, and Growth", *Journal of Economic Growth,* 2, pp. 1-26.

Baumol, W. J. (1986): "Productivity Growth, Convergence, and Welfare", *American Economic Review,* 76, pp. 1072-85.

Benhabib, J, and M. Spiegel (2002): "Human Capital and Technology Diffusion", Unpublished, NYU.

Bernanke, B., and M. Gertler (1989): "Agency Costs, Net Worth, and Business Fluctuations", *American Economic Review,* 79, pp. 14-31.

Blundell, R, Griffith, R, and J. Van Reenen (1999): "Market Share, Market Value and Innovation in a Panel of British Manufacturing Firms", *Review of Economic Studies,* 66, 529-554.

Boone, J (2000): "Measuring Product Market Competition", CEPR Discussion Paper 2636.

Coe, D. T., and E. Helpman (1995): "International R&D Spillovers", *European Economic Review,* 39, pp. 859-87.

Coe, D. T., E. Helpman, and A. W. Hoffmaister (1996): "North-South R& D Spillovers", *Economic Journal,* 107, pp. 134-49.

Cohen, W, R. Levin (1989): "Empirical Studies of Innovation and Market Structure", Chapter 18 of R. Schalensee and R. Willig, *Handbook of Industrial Organization,* Elsevier.

Cohen, W. M., and D.A. Levinthal (1989): "Innovation and Learning: The Two Faces of R&D", *Economic Journal,* 99, pp. 569-96.

Corriveau, L (1991): "Entrepreneurs, Growth and Cycles," PhD Dissertation, University of Western Ontario.

Dinopoulos, E, and P. Thompson (1998): "Schumpeterian Growth without Scale Effects", *Journal of Economic Growth,* 3, pp. 313-35.

Djankov, S, La Porta, R, Lopez-de-Silanes, F, and A Shleifer (2002): "The Regulation of Entry," *Quarterly Journal of Economics*, CXVII.

Durlauf, S, and P. Johnson (1995): "Multiple Regimes and Cross-Country Growth Behavior", *Journal of Applied Econometrics,* 10, pp. 365-84.

Eaton, J, and S. Kortum (1996): "Trade in Ideas: Patenting and Productivity in the OECD", *Journal of International Economics,* 40, pp. 251-78.

Eaton, J, and S. Kortum (2001): "Technology, Trade, and Growth: A Unified Framework", *European Economic Review,* 45, pp. 742-55.

Evans, P (1996): "Using Cross-Country Variances to Evaluate Growth Theories", *Journal of Economic Dynamics and Control,* 20, pp. 1027-49.

Evenson, R. E., and L. E. Westphal (1995): "Technological Change and Technology Strategy", In *Handbook of Development Economics,vol 3A*, edited by T. N. Srinivasan and Jere Behrman, 2209-99. Amsterdam: Elsevier.

Frankel, J, and D. Romer (1999): "Does Trade Cause Growth?", *American Economic Review,* 89, 379-399.

Galor, O, and J. Zeira (1993): "Income Distribution and Macroeconomics", *Review of Economic Studies,* 60, 35-52.

Geroski, P (1995): *Market Structure, Corporate Performance and Innovative Activity*, Oxford University Press.

Gerschenkron, A (1952): "Economic Backwardness in Historical Perspective", In *The Progress of Underdeveloped Areas*, edited by Bert F. Hoselitz. Chicago: University of Chicago Press.

Griffith, R, S. Redding, and J. Van Reenen (2001): "Mapping the Two Faces of R&D: Productivity Growth in a Panel of OECD Industries", unpublished.

Grossman, G, and E. Helpman (1991): *Innovation and Growth in the Global Economy,* MIT Press.

Grossman, S, and O. Hart (1986): "The Costs and Benefits of Ownership: A Theory of Lateral and Vertical Integration", *Journal of Political Economy, 94, 691-719.*

Ha, J, and P. Howitt (2004): "Accounting for Trends in Productivity and R&D: A Schumpeterian Critique of Semi-Endogenous Growth Theory", unpublished..

Howitt, P (1999): "Steady Endogenous Growth with Population and R&D Inputs Growing", *Journal of Political Economy,* 107, pp. 715-30.

Howitt, P (2000): "Endogenous Growth and Cross-Country Income Differences", *American Economic Review,* 90, 829-46.

Howitt, P, and D. Mayer-Foulkes (2002): "R&D, Implementation and Stagnation: A Schumpeterian Theory of Convergence Clubs", NBER Working Paper 9104.

Jones, C (1995): "R&D-Based Models of Economic Growth", *Journal of Political Economy,* 103, 759-84.

Jones, C (1999): "Growth: With or Without Scale Effects?" *American Economic Review, Papers and Proceedings,* 89, 139-44..

Keller, W (2002): "Technology Diffusion and the World Distribution of Income: The Role of Geography, Language, and Trade", Unpublished, University of Texas.

van de Klundert, T, and S. Smulders (1997): "Growth, Competition, and Welfare", *Scandinavian Journal of Economics,* 99, pp. 99-118.

La Porta, R, F. Lopez-de-Silanes, A. Shleifer, and R. W. Vishny (1998): "Law and Finance", *Journal of Political Economy,* 106, pp. 1113-55.

Levine, R, N. Loayza, and T. Beck (2000): "Financial Intermediation and Growth: Causality and Causes", *Journal of Monetary Economics,* 46, pp. 31-77.

Lucas, R (1988): "On the Mechanics of Economic Development", *Journal of Monetary Economics, 22, 3-42.*

Maddison, A (2001): *The World Economy: A Millennial Perspective.* Development Centre Studies. Paris: OECD.

Mankiw, G, Romer, D and D. Weil (1992): "A Contribution to the Empirics of Economic Growth", *Quarterly Journal of Economics,* 107, pp. 407-37.

Mayer-Foulkes, D (2002): "Global Divergence", Documento de Trabajo del CIDE, SDTE 250, División de Economía.

Mayer-Foulkes, D (2003): "Convergence Clubs in Cross-Country Life Expectancy Dynamics", In *Perspectives on Growth and Poverty*, edited by Rolph van der Hoeven and Anthony F. Shorrocks, 144-71. Tokyo: United Nations University Press.

Nelson, R, and E. Phelps (1966): "Investment in Humans, Technological Diffusion, and Economic Growth", *American Economic Review, 61, 69-75.*

Nickell, S (1996): "Competition and Corporate Performance", *Journal of Political Economy,* 104, 724-746

Peretto, P. F. (1998): "Technological Change, Market Rivalry, and the Evolution of the Capitalist Engine of Growth", *Journal of Economic Growth,* 3, pp. 53-80.

Piketty, T, Postel-Vinay, G, and J-L Rosenthal (2003): " Wealth Concentration in a Developing Economy: Paris and France, 1807-1994", mimeo EHESS (Paris).

Pritchett, L (1997): "Divergence, Big-Time", *Journal of Economic Perspectives,* 11, pp. 3-17.

Quah, D (1993): "Empirical Cross-Section Dynamics in Economic Growth", *European Economic Review,* 37, pp. 426-34.

Quah, D (1996): "Convergence Empirics Across Economies with (Some) Capital Mobility", *Journal of Economic Growth,* 1,: 95-124.

Quah, D (1997): "Empirics for Growth and Distribution: Stratification, Polarization, and Convergence Clubs", *Journal of Economic Growth,* 2, pp. 27-59.

Romer, P (1990): "Endogenous Technical Change", *Journal of Political Economy,*

Savvides, A, and M. Zachariadis (2004): "International Technology Diffusion and the Growth of TFP in the Manufacturing Sector of Developing Economies", *Review of Development Economics*, forthcoming.

Scherer, F (1967): "Market Structure and the Employment of Scientists and Engineers", *American Economic Review,* 57, 524-531.

Segerstrom, P, Anant, T, and E. Dinopoulos (1990): "A Schumpeterian Model of the Product Life Cycle", *American Economic Review, 80, 1077-1092*

Tirole, J (1988): *The Theory of Industrial Organization.* Cambridge, MA: MIT Press.

Young, A (1998): "Growth without Scale Effects", *Journal of Political Economy*, 106, 41-63.

Zachariadis, M (2003): "R&D, Innovation, and Technological Progress: A Test of the Schumpeterian Framework without Scale Effects", *Canadian Journal of Economics,* 36, pp. 566-86.

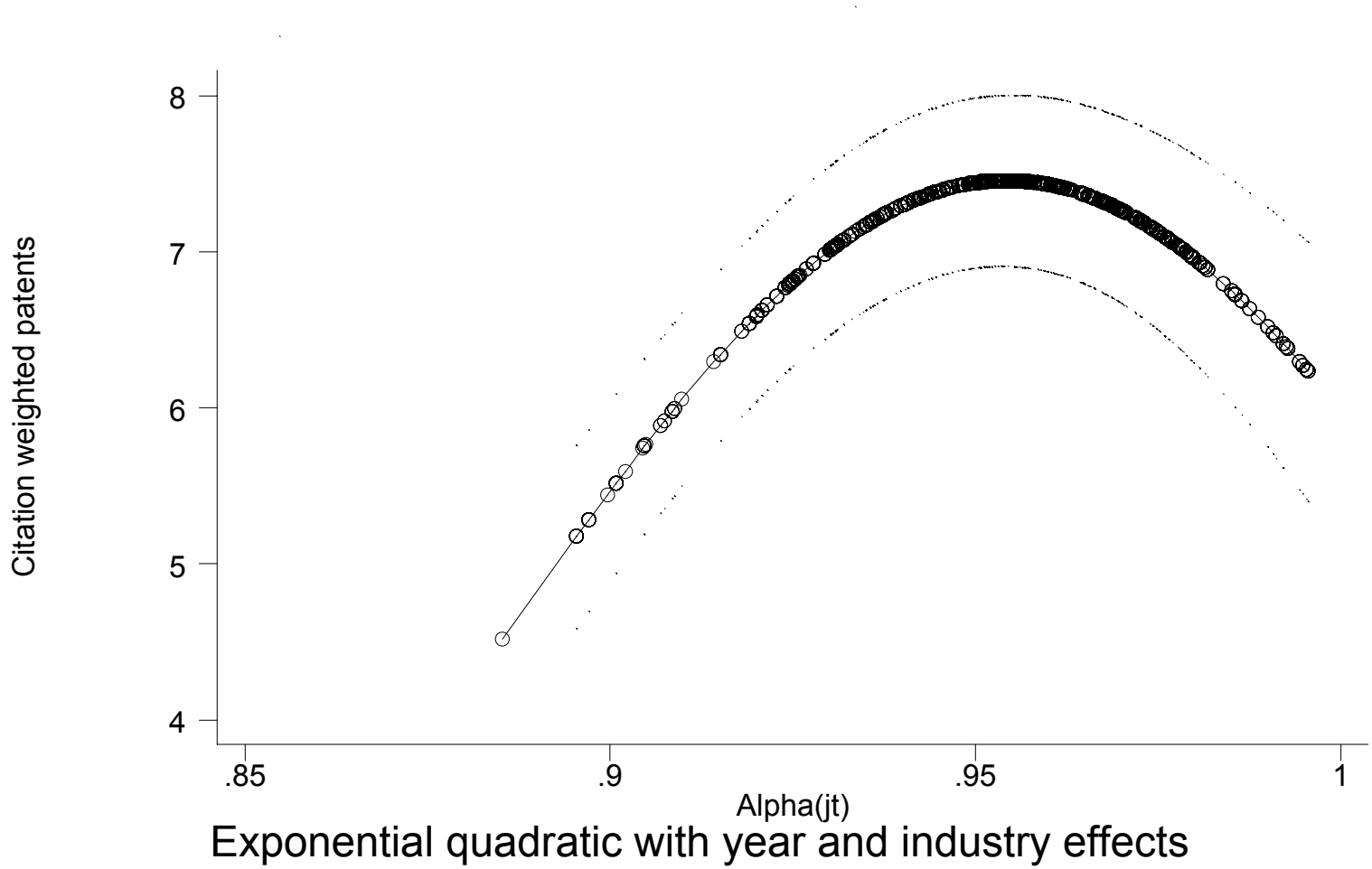
Table 1: Growth, Financial Development, and Initial GDP Gap

Estimation of equation: $g - g_1 = \beta_0 + \beta_f F + \beta_y (y - y_1) + \beta_{fy} F (y-y_1) + \beta_x X$

Financial development *(F)*	Private Credit			Liquid Liabilities			Bank Assets		
Conditioning set (X)	Empty	Policy[a]	Full[b]	Empty	Policy[a]	Full[b]	Empty	Policy[a]	Full[b]
Coefficient estimates									
β_f	-0.015	-0.013	-0.016	-0.029	-0.030	-0.027	-0.019	-0.020	-0.022
	(-0.93)	(-0.68)	(-0.78)	(-1.04)	(-0.99)	(-0.90)	(-1.07)	(-1.03)	(-1.12)
β_y	1.507***	1.193*	1.131	2.648***	2.388**	2.384**	1.891***	1.335*	1.365
	(3.14)	(1.86)	(1.49)	(3.12)	(2.39)	(2.11)	(3.57)	(1.93)	(1.66)
β_{fy}	-0.061***	-0.063***	-0.063***	-0.076***	-0.077***	-0.073***	-0.081***	-0.081***	-0.081***
	(-5.35)	(-5.10)	(-4.62)	(-3.68)	(-3.81)	(-3.55)	(-5.07)	(-4.85)	(-4.46)
sample size	71	63	63	71	63	63	71	63	63

Notes: The dependent variable $g-g_1$ is the average growth rate of per-capita real GDP relative to the US, 1960-95. F is average Financial Development 1960-95 using 3 alternative measures: Private Credit is the value of credits by financial intermediaries to the private sector, divided by GDP, Liquid Liabilities is currency plus demand and interest-bearing liabilities of banks and non-bank financial intermediaries, divided by GDP, and Bank Assets is the ratio of all credits by banks to GDP. $y - y_1$ is the log of per-capita GDP in 1960 relative to the United States. [a]The Policy conditioning set includes average years of schooling in 1960, government size, inflation, the black market premium and openness to trade. [b]The Full conditioning set includes the policy set plus indicators of revolutions and coups, political assassinations and ethnic diversity. Estimation is by IV using L (legal origins) and $L\ (y-y_1)$ as instruments for F and $F\ (y-y_1)$. The numbers in parentheses are t-statistics. Significance at the 1%, 5% and 10% levels is denoted by ***, ** and * respectively.

Figure 1: Innovation and Product Market Competition



Chapter 3

HORIZONTAL INNOVATION IN THE THEORY OF GROWTH AND DEVELOPMENT

GINO GANCIA
CREI and Universitat Pompeu Fabra

FABRIZIO ZILIBOTTI
Institute for International Economic Studies, Stockholm University

Contents

Handbook of Economic Growth, Volume 1A. Edited by Philippe Aghion and Steven N. Durlauf

DOI: 10.1016/S1574-0684(05)01003-8

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 2
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

Abstract

We analyze recent contributions to growth theory based on the model of expanding variety of Romer [Romer, P. (1990). "Endogenous technological change". Journal of Political Economy 98, 71–102]. In the first part, we present different versions of the benchmark linear model with imperfect competition. These include the "lab-equipment" model, "labor-for-intermediates" and "directed technical change". We review applications of the expanding variety framework to the analysis of international technology diffusion, trade, cross-country productivity differences, financial development and fluctuations. In many such applications, a key role is played by complementarities in the process of innovation.

Keywords

appropriate technology, complementarity, cycles, convergence, directed technical change, endogenous growth, expanding variety, financial development, imperfect competition, integration, innovation, intellectual property rights, imitation, knowledge, learning, patents, technical change, trade, traps

JEL classification: D92, E32, F12, F15, F43, G22, O11, O16, O31, O33, O41, O47

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 3
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

1. Introduction

Endogenous growth theory formalizes the role of technical progress in explaining modern economic growth. Although this is a relatively recent development, many of its ideas were already stressed by authors such as Kuznets, Griliches, Schmookler, Rosenberg and Schumpeter. During the 1950s and 1960s, mainstream economics was dominated by the one-sector neoclassical growth model of Solow (1956) and Swan (1956), whose main focus was on capital accumulation. The model postulated the existence of an aggregate production function featuring constant returns to scale and returns to each input falling asymptotically to zero; given that some inputs cannot be accumulated, the model could not generate sustained growth unless technology was assumed to improve exogenously. This simple treatment of technology as exogenous was considered as unsatisfactory for two main reasons: first, by placing the source of sustained growth outside the model, the theory could not explain the determinants of long-run economic performance and second, empirical evidence pointed out that technical progress often depends on deliberate economic decisions.

The first attempts to endogenize the rate of technical change addressed the first, but not the second, problem. Assuming technical progress to be an unintentional by-product of the introduction of new capital goods through a process named "learning-by-doing", Arrow (1962) was able to generate sustained growth at a rate that depended on investment decisions. Attempts at explicitly modeling investment in innovation faced another difficulty. A replication argument suggests that, for a given state of technology, production functions should exhibit constant returns to scale. If technical progress is considered as an additional input, however, the technology features increasing returns to scale and inputs cannot be paid their marginal product. Models of learning-by-doing avoided the problem by assuming that increasing returns were external to firms, thereby preserving perfect competition. However, this approach is not viable once investment in technology is recognized as intentional. The solution was to follow the view of Schumpeter (1942), that new technologies provide market power and that investment in innovation is motivated by the prospect of future profits. In this spirit, Shell (1973) studied the case of a single monopolist investing in technical change and Nordhaus (1969a) wrote a growth model with patents, monopoly power and many firms. In neither case did the equilibrium feature sustained growth.[1]

A tractable model of imperfect competition under general equilibrium was not available until the analysis of monopolistic competition in consumption goods by Dixit and Stiglitz (1977), later extended to differentiated inputs in production by Ethier (1982). These models also showed how increasing returns could arise from an expansion in the number of varieties of producer and consumer goods, an idea that is at the core of the models studied in this chapter. The first dynamic models of economic growth with monopolistic competition and innovation motivated by profits were built by Judd (1985)

[1] See Levhari and Sheshinski (1969) on necessary and sufficient conditions for the existence of steady-state growth in the presence of increasing returns to scale.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 4
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

and Grossman and Helpman (1989). Yet, these authors were interested in aspects other than endogenous growth and none of their models featured long-run growth. Romer (1987), who formalized an old idea of Young (1928), was the first to show that models of monopolistic competition could generate long-run growth through the increased specialization of labor across an increasing range of activities. The final step was taken in Romer (1990), which assumed that inventing new goods is a deliberate costly activity and that monopoly profits, granted to innovators by patents, motivate discoveries. Since then, the basic model of endogenous growth with an expanding variety of products has been extended in many direction.

The distinctive feature of the models discussed in this chapter is "horizontal innovation": a discovery consists of the technical knowledge required to manufacture a new good that does not displace existing ones. Therefore, innovation takes the form of an expansion in the variety of available products. The underlying assumption is that the availability of more goods, either for final consumption or as intermediate inputs, raises the material well-being of people. This can occur through various channels. Consumers may value variety per se. For example, having a TV set and a Hi-Fi yields more utility than having two units of any one of them. Productivity in manufacturing may increase with the availability of a larger set of intermediate tools, such as hammers, trucks, computers and so on. Similarly, specialization of labor across an increasing variety of activities, as in the celebrated Adam Smith example of the pin factory, can make aggregate production more efficient. The main alternative approach is to model innovation as quality improvements on a given array of products ("vertical innovation"), so that technical progress makes existing products obsolete. This process of "creative destruction" was emphasized by Schumpeter and has been formalized in Aghion and Howitt (1992), Grossman and Helpman (1991a, 1991b) and Segerstrom, Anant and Dinopoulos (1990). The two approaches naturally complement each other. The main advantage of models with horizontal innovation lies in their analytical tractability, making them powerful tools for addressing a wide range of questions. However, because of their simplistic view on the interaction between innovators, these models are less suited to studying the effects of competition between " leaders" and "follower" on the growth process.

Section 1 of this chapter describes a simplified version of Romer (1990) and some extensions used in the literature. The model exhibits increasing returns to scale and steady-state endogenous growth in output per capita and the stock of knowledge. The key feature of the theory is the emphasis on investments in technical knowledge as the determinant of long-run economic growth. Ideas and technological improvements differ from other physical assets, because they entail important public good elements. Inventing new technology is typically costly, while reproducing ideas is relatively inexpensive. Therefore, technical knowledge is described as a non-rival good. Nevertheless, firms are willing to invest in innovation because there exists a system of intellectual property rights (patents) guaranteeing innovators monopoly power over the production and sales of particular goods.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 5
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

Growth models with an expanding variety of products are a natural dynamic counterpart to trade models based on increasing returns and product differentiation. As such, they offer a simple framework for studying the effects of market integration on growth and other issues in dynamic trade theory. This is the subject of Section 2, which shows how trade integration can produce both static gains, by providing access to foreign varieties, and dynamic gains, by raising the rate at which new goods are introduced. Product-cycle trade and imitation are also considered.

In many instances, technical progress may be non-neutral towards different factors or sectors. This possibility is considered in Section 3, where biased technical change is incorporated in the basic growth model. By introducing several factors and sectors, the economic incentives to develop technologies complementing a specific factor, such as skilled workers, can be studied. These incentives critically depend on the definition of property rights over the production of new ideas. The high variability in the effectiveness of patent laws across countries has important bearings on the form of technical progress. In particular, governments in less developed countries may have an incentive not to enforce intellectual property rights in order to speed up the process of technology adoption. However, the undesired side effect of free-riding is that innovators in industrialized countries lose incentives to create improvements that are most useful in developing countries, but of limited application in industrialized markets.

Section 4 introduces complementarity in innovation. While innovation has no effect on the profitability of existing intermediate firms in the benchmark model, in reality new technologies can substitute or complement existing technologies. Innovation may cause technological obsolescence of previous technologies, as emphasized by Schumpeterian models. In other cases, new technologies complement rather than substitute the old ones. For instance, the market for a particular technology tends to be small at the time of its introduction, but grows as new compatible applications are developed. This complementarity in innovation can lead to multiple equilibria and poverty traps.

Complementarities in the growth process may also arise from financial markets, as suggested in Section 5. The progressive endogenous enrichment of asset markets, associated with the development of new intermediate industries, may improve the diversification opportunities available to investors. This, in turn, makes savers more prepared to invest in high-productivity risky industries, thereby fostering further industrial and financial development. As a result, countries at early stages of development go through periods of slow and highly volatile growth, eventually followed by a take-off with financial deepening and steady growth.

Finally, Section 6 shows how models with technological complementarities can generate rich long-run dynamics, including endogenous fluctuations between periods of high and low growth. Cycles in innovation and growth can either be due to expectational indeterminacy, or the deterministic dynamics of two-sector models with an endogenous market structure.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 6
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

2. Growth with expanding variety

In this section, we present the benchmark model of endogenous growth with expanding variety, and some extensions that will be developed in the following sections.

2.1. The benchmark model

The benchmark model is a simplified version of Romer (1990), where, for simplicity, we abstract from investments in physical capital. The economy is populated by infinitely lived agents who derive utility from consumption and supply inelastic labor. The population is constant, and equal to L. Agents' preferences are represented by an isoelastic utility function:

$$U = \int_0^\infty \mathrm{e}^{-\rho t} \frac{C_t^{1-\theta} - 1}{1-\theta} \,\mathrm{d}t. \tag{1}$$

The representative household sets a consumption plan to maximize utility, subject to an intertemporal budget constraint and a No-Ponzi game condition. The consumption plan satisfies a standard Euler equation:

$$\dot{C}_t = \frac{r_t - \rho}{\theta} \cdot C_t. \tag{2}$$

There is no physical capital, and savings are used to finance innovative investments.

The production side of the economy consists of two sectors of activity: a competitive sector producing a homogeneous final good, and a non-competitive sector producing differentiated intermediate goods. The final-good sector employs labor and a set of intermediate goods as inputs. The technology for producing final goods is represented by the following production function:

$$Y_t = L_{y,t}^{1-\alpha} \int_0^{A_t} x_{j,t}^{\alpha} \,\mathrm{d}j, \tag{3}$$

where x_j is the quantity of the intermediate good j, A_t is the measure of intermediate goods available at t, L_y is labor and $\alpha \in (0, 1)$. This specification follows Spence (1976), Dixit and Stiglitz (1977) and Ethier (1982). It describes different inputs as imperfect substitutes, which symmetrically enter the production function, implying that no intermediate good is intrinsically better or worse than any other, irrespective of the time of introduction. The marginal product of each input is decreasing, and independent of the measure of intermediate goods, A_t.

The intermediate good sector consists of monopolistically competitive firms, each producing a differentiated variety j. Technology is symmetric across varieties: the production of one unit of intermediate good requires one unit of final good, assumed to be the numeraire.[2] In addition, each intermediate producer is subject to a sunk cost to

[2] In Romer (1990), the variable input is physical capital, and the economy has two state variables, i.e., physical capital and knowledge.

design a new intermediate input variety. New designs are produced instantaneously and with no uncertainty. The innovating firm can patent the design, and acquire a perpetual monopoly power over the production of the corresponding input.

In the absence of intellectual property rights, free-riding would prevent any innovative activity. If firms could costlessly copy the design, competition would drive ex-post rents to zero. Then, no firms would have an incentive, ex-ante, to pay a sunk cost to design a new input.

The research activity only uses labor. An important assumption is that innovation generates an intertemporal externality. In particular, the design of a (unit measure of) new intermediate good requires a labor input equal to $1/(\delta A_t)$. The assumption that labor productivity increases with the stock of knowledge, A_t, can be rationalized by the idea of researchers benefiting from accessing the stock of applications for patents, thereby obtaining inspiration for new designs.

The law of motion of technical knowledge can be written as:

$$\dot{A}_t = \delta A_t L_{x,t}, \tag{4}$$

where δ is a parameter and L_x denotes the aggregate employment in research. The rate of technological change is a linear function of total employment in research.[3] Finally, feasibility requires that $L \geqslant L_{x,t} + L_{y,t}$.

First, we characterize the equilibrium in the final good sector. Let w denote the wage, and p_j be the price of the j'th variety of intermediate input. The price of the final product is the numeraire. The representative firm in the competitive final sector takes prices as parametric and chooses production and technology so as to maximizes profit, given by:

$$\pi_t^Y = L_{y,t}^{1-\alpha} \int_0^{A_t} {x_{j,t}}^{\alpha} \,\mathrm{d}j - w_t L_{y,t} - \int_0^{A_t} p_{j,t} x_{j,t} \,\mathrm{d}j. \tag{5}$$

The first-order conditions yield the following factor demands:

$$p_{j,t} = \alpha L_{y,t}^{1-\alpha} {x_{j,t}}^{\alpha-1} \quad \forall j \in [0, A_t] \tag{6}$$

and

$$w_{y,t} = (1-\alpha) L_{y,t}^{-\alpha} \int_0^{A_t} {x_{j,t}}^{\alpha} \,\mathrm{d}j. \tag{7}$$

[3] Jones (1995) generalizes this technology and lets

$$\dot{A}_t = \delta A_t^{\gamma_a} L_{x,t}^{(1-\gamma_L)},$$

where $\gamma_A \leqslant 1$ is a positive externality through the stock of knowledge and γ_L is a negative externality that can be interpreted as coming from the duplication of research effort. Assuming $\gamma_A < 1$ leads to qualitative differences in the prediction of the model. In particular, the specification where $\gamma_A = 1$ and $\gamma_L = 0$, which is the model discussed here, generates scale effects. See further discussion later in this chapter and, especially, in Chapter ? of this Handbook.

Next, consider the problem of intermediate producers. A firm owning a patent sets its production level so as to maximize the profit, subject to the demand function (6). The profit of the firm producing the jth variety is $\pi_{j,t} = p_{j,t}x_{j,t} - x_{j,t}$. The optimal quantity and price set by the monopolist are

$$x_{j,t} = x_t = \alpha^{2/(1-\alpha)}L_{y,t} \quad \text{and} \quad p_{j,t} = p = 1/\alpha, \tag{8}$$

respectively. Hence, the maximum profit for an intermediate producer is

$$\pi_{j,t} = \pi_t = (p-1)x_t = \frac{1-\alpha}{\alpha}\alpha^{2/(1-\alpha)}L_{y,t}. \tag{9}$$

Substitution of x_t into (7) yields the equilibrium wage as:

$$w_t = (1-\alpha)\alpha^{2\alpha/(1-\alpha)}A_t. \tag{10}$$

Next, we guess-and-verify the existence of a balanced growth (BG) equilibrium, such that consumption, production and technical knowledge grow at the same constant rate, γ, and the two sectors employ constant proportions of the workforce.[4] In BG, both the production and the profits of intermediate firms, as given by Equations (8) and (9), are constant over time and across industries. Thus, $x_t = x$ and $\pi_t = \pi$.

Free entry implies that the present discounted value (PDV) of profits from innovation cannot exceed the entry cost. By the Euler equation, (2), the interest rate is also constant in BG. Hence, the PDV of profits equals π/r. The entry cost is given by the wage paid to researchers, i.e., $w_t/(\delta A_t)$. Therefore, the free entry condition can be written as:

$$\frac{\pi}{r} \leqslant \frac{w_t}{\delta A_t}. \tag{11}$$

We can then use (9) and (10), and substitute the expressions of π and w_t into (11):

$$\frac{(\frac{1-\alpha}{\alpha})\alpha^{2/(1-\alpha)}L_y}{r} \leqslant \frac{(1-\alpha)\alpha^{2\alpha/(1-\alpha)}}{\delta}. \tag{12}$$

The right-hand side expression is the marginal cost of innovation, independent of A_t, due to the cancellation of two opposite effects. On the one hand, labor productivity and, hence, the equilibrium wage grow linearly with A_t. On the other hand, the productivity of researchers increases with A_t, due to the intertemporal knowledge spillover. Thus, the unit cost of innovation is constant over time. Note that, without the externality, the cost of innovation would grow over time, and technical progress and growth would come to a halt, like in the neoclassical model.

For innovation to be positive, (12) must hold with equality. We can use (i) the resource constraint, implying that $L_y = L - L_x$, and (ii) the fact that, from (4) and BG, $L_x = \gamma/\delta$, to express (12) as a relationship between the interest rate and the growth rate:

$$r = \alpha(\delta L - \gamma). \tag{13}$$

[4] The equilibrium that we characterized can be proved to be unique. Moreover, the version of Romer's model described here features no transitional dynamics, as in AK models [Rebelo (1991)].

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 9
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

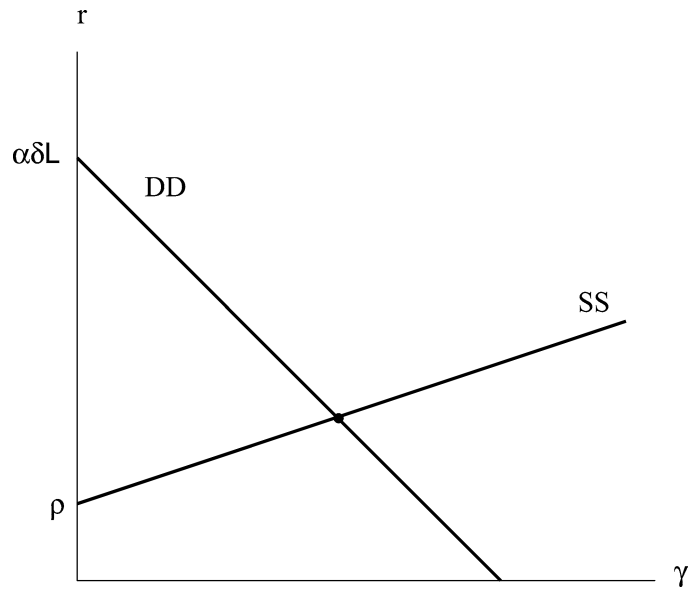


Figure 1.

Equation (13) describes the equilibrium condition on the production side of the economy: the higher is the interest rate that firms must pay to finance innovation expenditure, the lower is employment in research and growth.

Finally, the consumption Euler equation, (2), given BG, yields:

$$r = \rho + \theta\gamma, \tag{14}$$

which is the usual positive relation between interest rate and growth. Figure 1 plots the linear equations (13) and (14), which characterize the equilibrium. The two equations correspond, respectively, to the DD (demand for funds) and SS (supply of savings) linear schedules.

An interior solution exists if and only if $\alpha\delta L > \rho$. When this condition fails to be satisfied, all workers are employed in the production of consumption goods. When it is positive, the equilibrium growth rate is

$$\gamma = \frac{\delta\alpha L - \rho}{\alpha + \theta}, \tag{15}$$

showing that the growth rate is increasing in the productivity of the research sector (δ), the size of the labor force (L) and the intertemporal elasticity of substitution of consumption ($1/\theta$), while it is decreasing in the elasticity of final output to labor, ($1 - \alpha$), and the discount rate.

The trade-off between final production (consumption), on the one hand, and innovation and growth, on the other hand, can be shown by substituting the equilibrium

expression of x into the aggregate production function, (3). This yields:

$$Y_t = \alpha^{2\alpha/(1-\alpha)} L_y A_t = \alpha^{2\alpha/(1-\alpha)} (L - \gamma/\delta) A_t. \tag{16}$$

The decentralized equilibrium is inefficient for two reasons:[5]

1. Intermediate firms exert monopoly power, and charge a price in excess of the marginal cost of production. This leads to an underproduction of each variety of intermediate goods.
2. the accumulation of ideas produces externalities not internalized in the laissez-faire economy. Innovating firms compare the private cost of innovation, $w_t/(\delta A_t)$, with the present discounted value of profits, π/r. However, they ignore the spillover on the future productivity of innovation.

Contrary to Schumpeterian models, innovation does not cause "creative destruction", i.e., no rent is reduced by the entry of new firms. As a result, growth is always suboptimally low in the laissez-faire equilibrium. Policies aimed at increasing research activities (e.g., through subsidies to R&D or intermediate production) are both growth- and welfare-enhancing. This result is not robust, however. Benassy (1998) shows that in a model where the return to specialization is allowed to vary and does not depend on firms' market power (α), research and growth in the laissez-faire equilibrium may be suboptimally too high.

2.2. Two variations of the benchmark model: "lab-equipment" and "labor-for intermediates"

We now consider two alternative specifications of the model that have been used in the literature, and that will be discussed in the following sections. The first specification is the so-called "lab-equipment" model, where the research activity uses final output instead of labor as a productive input.[6] More formally, Equation (4) is replaced by the condition $\dot{A}_t = Y_x/\mu$, where Y_x denotes the units of final output devoted to research (hence, consumption is $C = Y - Ax - Y_x$) and μ the output cost per unit of innovation. In the lab-equipment model, there is no research spillover of the type discussed in the benchmark model. Labor is entirely allocated to final production ($L_y = L$), and the free-entry condition (12) is replaced by

$$\frac{\frac{1-\alpha}{\alpha}\alpha^{2/(1-\alpha)} L}{r} \leqslant \mu. \tag{17}$$

[5] There is an additional reason why, in general, models with a Dixit–Stiglitz technology can generate inefficient allocations in laissez-faire, namely that the range of intermediate goods produced is endogenous. The standard assumption of complete markets is violated in Dixit–Stiglitz models, because there is no market price for the goods not produced. This issue is discussed in Matsuyama (1995, 1997). A dynamic example of such a failure is provided by the model of Acemoglu and Zilibotti (1997), which is discussed in detail in Section 6.

[6] The "lab-equipment" model was first introduced by Rivera-Batiz and Romer (1991a); see also Barro and Sala-i-Martin (1995).

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 11
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

Hence, using the Euler condition, (14), we obtain the following equilibrium growth rate:

$$\gamma = \frac{(1-\alpha)\alpha^{(1+\alpha)/(1-\alpha)}L/\mu - \rho}{\theta}.$$

Sustained growth is attained by allocating a constant share of production to finance the research activity.

The second specification assumes that labor is not used in final production, but is used (instead of final output) as the unique input in the intermediate goods production.[7] More formally, the final production technology is

$$Y_t = Z^{1-\alpha}\int_0^{A_t} x_{j,t}^{\alpha}\,\mathrm{d}j, \tag{18}$$

where Z is a fixed factor (e.g., land) that is typically normalized to unity and ignored. In this model, $1/A_t$ units of labor are required to produce one unit of any intermediate input, with constant marginal costs. Therefore, in this version of the model, innovation generates a spillover on the productivity of both research and intermediate production.[8] We refer to this version as the "labor-for-intermediates" model.

It immediately follows that, in equilibrium, the production of each intermediate firm equals $x = L - L_x$. The price of intermediates is once more a mark-up over the marginal cost, $p_t = w_t/(\alpha A_t)$. In a BG equilibrium, wages and technology grow at the same rate, hence their ratio is constant. Let $\omega \equiv (w_t/A_t)$. The maximum profit is, then:

$$\pi = \left(\frac{1-\alpha}{\alpha}\right)\omega x = \frac{1-\alpha}{\alpha}\omega(L - L_x).$$

The free entry condition can be expressed as:

$$\frac{1-\alpha}{r\alpha}(L - L_x) \leqslant \frac{1}{\delta},$$

hence,

$$\gamma = \left(\frac{1-\alpha}{\alpha}L\delta - \rho\right)\Big/\left(\frac{1-\alpha}{\alpha} + \theta\right).$$

Clearly, both the "lab-equipment" and "labor-for-intermediates" model yield solutions qualitatively similar to that of the benchmark model.

[7] We follow the specification used by Young (1993). A related approach, treating the variety of inputs as consumption goods produced with labor, is examined in Grossman and Helpman (1991a, 1991b).

[8] The spillover on the productivity of intermediate production is not necessary to have endogenous growth. Without it, an equilibrium can be found in which production of each intermediate falls as A grows: $\gamma_A = -\gamma_x$. In this case, employment in production, Ax, is constant and the growth rate of Y is $(1-\alpha)\gamma_A$.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 12
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

2.3. Limited patent protection

In this section, we discuss the effects of limited patent protection. For simplicity, we focus on the lab-equipment version discussed in the previous section. The expectation of monopoly profits provides the basic incentive motivating investment in innovation; at the same time, monopoly rights introduce a distortion in the economy that raises prices above marginal costs and causes the underprovision of goods. Since the growth rate of knowledge in the typical decentralized equilibrium is below the social optimum, the presence of monopoly power poses a trade-off between dynamic and static efficiency, leading to the question, first studied by Nordhaus (1969a, 1969b), of whether there exists an optimal level of protection of monopoly rights. In the basic model, we assumed the monopoly power of innovators to last forever. Now, we study how the main results change when agents cannot be perfectly excluded from using advances discovered by others. A tractable way of doing this is to assume monopoly power to be eroded at a constant rate, so that in every instant, a fraction m of the monopolized goods becomes competitive.[9] Then, for a given range of varieties in the economy, A_t, the number of "imitated" intermediates that have become competitive, A_t^*, follows the law of motion:

$$\dot{A}_t^* = m\big(A_t - A_t^*\big). \tag{19}$$

Stronger patent protection can be considered as a reduction in the imitation rate m. Note that the model now has two state variables, A_t and A_t^*, and will exhibit transitional dynamics. In general, from any starting point, the ratio A_t^*/A_t will converge to the steady-state level:[10]

$$\frac{A^*}{A} = \frac{m}{\gamma + m}, \tag{20}$$

where $\gamma \equiv \dot{A}/A$.

Once a product is imitated, the monopoly power of the original producer is lost and its prices is driven down to the marginal cost by competition. Thus, at each point in time, intermediates still produced by monopolists are sold as before at the markup price $1/\alpha$, while for the others, the competitive price is one. Substituting prices into demand functions yields the quantity of each intermediate sold in equilibrium:

$$\begin{aligned} x_j &= \alpha^{1/(1-\alpha)} L \equiv x^* \quad \text{for } j \in \big(0, A_t^*\big), \\ x_j &= \alpha^{2/(1-\alpha)} L \equiv x \quad \text{for } j \in \big(A_t^*, A_t\big). \end{aligned} \tag{21}$$

Note that $x^* > x$, because the monopolized goods have a higher price.

[9] A growth model with limited patent life is developed by Judd (1985). Here, we follow Barro and Sala-i-Martin (1995). An alternative way of introducing limited patent protection is to assume monopolies to have a deterministic lifetime T. In this case, the PDV of an innovation is $(1-\mathrm{e}^{-rT})\pi/r$ (assuming balanced growth).

[10] This can be seen imposing $\dot{A}^*/A^* = \gamma$ in (19).

Free entry requires the PDV of profits generated by an innovation, V, to equal its cost μ. Along the balanced growth path, where the interest rate is constant, arbitrage in asset markets requires the instantaneous return to innovation, π/μ, to equal the real interest rate adjusted for imitation risk: $r + m$.[11] Since prices and quantities of the monopolized goods are identical to those in the basic model, π is not affected by imitation. Imitation only affects the duration of the profit flow, which is reflected in the effective interest rate. Therefore, limiting patent lives introduces a new inefficiency: although the benefit from a discovery is permanent for the economy, the reward for the innovator is now only temporary. Using the Euler equation for consumption growth, $\gamma = (r - \rho)/\theta$, and the adjusted interest rate in (17), we get the growth rate of the economy:

$$\gamma = \frac{1}{\theta}\left[(1-\alpha)\alpha^{(1+\alpha)(1-\alpha)}\frac{L}{\mu} - m - \rho\right].$$

As expected, the growth rate is decreasing in the imitation rate, as the limited duration of the monopoly effectively reduces the private value of an innovation. If we were concerned about long-run growth only, it would then be clear that patents should always be fully and eternally protected. However, for a given level of technology, A_t, output is higher the shorter is the patent duration (higher m), as can be seen by substituting equilibrium quantities (21) and the ratio of imitated goods (20) in the production function (3):

$$Y_t = \alpha^{2\alpha/(1-\alpha)} A_t L\left[1 + \left(\frac{m}{\gamma + m}\right)\left(\alpha^{-\alpha/(1-\alpha)} - 1\right)\right].$$

Therefore, a reduction in the patent life entails a trade-off between an immediate consumption gain and future losses in terms of lower growth, and its quantitative analysis requires the calculation of welfare along the transition. Kwan and Lai (2003) perform such an analysis, both numerically and by linearizing the BG equilibrium in the neighborhood of the steady-state, and show the existence of an optimum patent life. They also provide a simple calibration, using US data on long-run growth, markups and plausible values for ρ and θ, to suggest that over-protection of patents is unlikely to happen, whereas the welfare cost of under-protection can be substantial.

Alternatively, the optimal patent length can be analytically derived in models with a simpler structure. For example, Grossman and Lai (2004) construct a modified version of the model described above, where they assume quasi-linear functions. They show the

[11] A simple way of seeing this is through the following argument. In a time interval $\mathrm{d}t$, the firm provides a profit stream $\pi \cdot \mathrm{d}t$, a capital gain of $\dot{V} \cdot \mathrm{d}t$ if not imitated and a capital loss V if imitated (as the value of the patent would drop to zero). In the limit $\mathrm{d}t \to 0$, the probability of being imitated in this time interval is $m \cdot \mathrm{d}t$ and the probability of not being imitated equals $(1 - m \cdot \mathrm{d}t)$. Therefore, the expected return for the firm is $\pi \cdot \mathrm{d}t + (1 - m \cdot \mathrm{d}t)\dot{V} \cdot \mathrm{d}t - mV \cdot \mathrm{d}t$. Selling the firm and investing the proceeds in the capital market would yield an interest payment of $rV \cdot \mathrm{d}t$. Arbitrage implies that the returns from these two forms of investment should be equal and in a steady state $\dot{V} = 0$, implying $\pi/V = r + m$.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 14
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

optimal patent length to be an increasing function of the useful life of a product, of consumers' patience and the ratio of consumers' and producers' surplus under monopoly to consumers' surplus under competition. In addition, they derive the optimal patent length for noncooperative trading countries and find that advanced economies with a higher innovative potential will, in general, grant longer patents. A similar point is made in Lai and Qiu (2003).

3. Trade, growth and imitation

Growth models with an expanding variety of products are a natural dynamic counterpart to the widely-used trade models based on increasing returns and product differentiation developed in the 1980s [e.g., Helpman and Krugman (1985)]. As such, they offer a simple framework for studying the effects of market integration on growth and other issues in dynamic trade theory. Quality-ladder models have also been proposed in this literature, but they are a less natural counterpart to the static new trade theory, as they do not focus on the number of varieties available in an economy and their growth rate. As we shall see, economic integration can provide both static gains, through the access to a wider range of goods, and dynamic gains, through an increase in the rate at which new varieties are introduced. However, the results may vary when integration is limited to commodity markets with no international diffusion of knowledge [Rivera-Batiz and Romer (1991a)] and when countries differ in their initial stock of knowledge [Devereux and Lapham (1994)].

Finally, the analysis in this section is extended to product-cycle trade: the introduction of new products in advanced countries and their subsequent imitation by less developed countries. An important result will be to show that, contrary to the closed economy case, imitation by less developed countries may spur innovation and growth [Helpman (1993)].

3.1. Scale effects, economic integration and trade

In this section, we use the benchmark model to discuss the effects of trade and integration. The model features scale effects. Take two identical countries with identical labor endowment, $L = L^*$. In isolation, both countries would grow at the same rate, as given by (15). But if they merge, the growth rate of the integrated country increases to:

$$\gamma^I = \frac{\delta\alpha(L + L^*) - \rho}{\alpha + \theta} = \frac{2\alpha\delta L - \rho}{\alpha + \theta}.$$

Therefore, the model predicts that economic integration boosts growth.

Integration, even if beneficial, may be difficult to achieve. However, in many instances, trade operates as a substitute for economic integration. Rivera-Batiz and Romer

(1991a) analyze under which condition trade would attain the same benefits as economic integration. To this aim, they consider two experiments:[12]

1. The economies can trade at no cost in goods and assets, but knowledge spillovers remain localized within national borders;
2. In addition, knowledge spillovers work across borders after trade.

In both cases, to simplify the analysis, the two economies are assumed to produce, before trade, disjoint subsets of intermediate goods. This assumptions avoids complications arising from trade turning monopolies into duopolies in those industries which exist in both countries. Clearly, after trade, there would be no incentive for overlap in innovation, and the importance of inputs that were historically produced in both countries would decline to zero over time.

We start from the case analyzed by Rivera-Batiz and Romer (1991a), where the two countries are perfectly identical before trade. Namely, $L = L^*$ and $A_0 = A_0^*$, where the star denotes the *foreign* economy, and time zero denotes the moment when trade starts. Since, in a BG equilibrium, $\gamma = \delta L_x$, trade can only affect growth via the split of the workforce between production and research. Such a split, however, is not affected by trade, for in the symmetric equilibrium, trade increases by the same proportion the productivity of workers in production and the profitability of research. Since both the cost and private benefit of innovation increase by the same factor, investments in innovation remain unchanged.

More formally, the after trade wage is

$$w_{\text{trade}} = (1-\alpha)L_y^{-\alpha}x^{\alpha}\big(A + A^*\big), \tag{22}$$

which is twice as large as in the pre-trade equilibrium since at the moment of trade liberalization, $A = A^*$. Higher labor costs are a disincentive to research. But trade also increases the market for intermediate goods. Each monopolist can now sell its product in two markets. Since the demand elasticity is the same in both markets, the monopoly price equals $1/\alpha$ in both markets. Thus, the after trade profit is

$$\pi_{\text{trade}} = (p-1)(x+x^*) = 2\frac{1-\alpha}{\alpha}\alpha^{2/(1-\alpha)}L_y.$$

The free-entry condition becomes, for both countries:

$$2\frac{\frac{1-\alpha}{\alpha}\alpha^{2/(1-\alpha)}L_y}{r} \leqslant 2\frac{(1-\alpha)\alpha^{2\alpha/(1-\alpha)}}{\delta}, \tag{23}$$

which, after simplifying, is identical to (12). Therefore, the split of the workforce between production and research remains unchanged, and trade has no permanent effects on growth. Opening up to free trade, however, induces a once-and-for-all gain: both output and consumption increase in both countries, similarly to an unexpected increase

[12] The original article considers two versions of the model, one using the benchmark set-up and the other using the "lab-equipment" version. For the sake of brevity, we restrict the attention to the first. Romer (1994) extends the analysis to the case when a tariff on imports is imposed.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 16
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

in the stock of knowledge, since final producers in both countries can use a larger set of intermediate goods.

This result is not robust to asymmetric initial conditions. Devereux and Lapham (1994) show that if, initially, the two countries have different productivity levels, trade leads to specialization and a rise in the world growth rate.[13] Consider the economies described above, but assume that $A_0 < A_0^*$. Recall that free-entry implies:

$$V \leqslant \frac{w}{\delta A} \quad \text{and} \quad V^* \leqslant \frac{w^*}{\delta A^*},$$

where V, V^* denote the PDV of profits for an intermediate firm located at home and abroad, respectively. First, trade in intermediate goods and free capital markets equalize the rate of return to both financial assets (r) and labor (w).[14] Second, monopoly profits are independent of firms' locations, thereby implying that the value of firms must be the same all over the world: $V = V^* = V^w$. Therefore, at the time of trade liberalization, we must have:

$$\frac{w_{\text{trade}}}{\delta A} > \frac{w_{\text{trade}}}{\delta A^*} \geqslant V^w,$$

implying that no innovation is carried out in equilibrium in the (home) country, starting from a lower productivity. Moreover, the productivity gap in R&D widens over time: indeed, trade forever eliminates the incentives to innovate in the initially poorer country.

In the richer (foreign) country, however, trade boosts innovation.[15] The value of foreign firms must satisfy the following Bellman equation:

$$rV^* = \dot{V}^* + \frac{1-\alpha}{\alpha}\alpha^{2/(1-\alpha)}\big(L_y^* + L\big),$$

where we note that $L_y = L$. The free-entry condition implies that:

$$V^* = \frac{(1-\alpha)\alpha^{2\alpha/(1-\alpha)}}{\delta}\frac{A^* + A}{A^*}.$$

Since knowledge only accumulates in the foreign country, the value of intermediate firms must decline over time, and in the long-run tend to its pre-trade value, i.e., $V^* = (1-\alpha)\alpha^{2\alpha/(1-\alpha)}/\delta$. Therefore, in the long run, the free-entry condition is

$$\frac{\frac{1-\alpha}{\alpha}\alpha^{2/(1-\alpha)}(L + L_y^*)}{r} \leqslant \frac{(1-\alpha)\alpha^{2\alpha/(1-\alpha)}}{\delta}. \tag{24}$$

[13] See also Rivera-Batiz and Romer (1991b) on the effects of trade restrictions with asymmetric countries.

[14] Recall Equation (22). The equalization of wages descends from a particular feature of the equilibrium, i.e., that the marginal product of labor is independent of the level of employment in production (since x is linear in L_y). This feature is not robust. If the production technology had land as an input, for instance, wages would not be equalized across countries; see Devereux and Lapham (1994) for an analysis of the more general case.

[15] Our discussion focuses on a world where no economy becomes fully specialized in research, since this seems to be the empirically plausible case.

Comparing (24) with (12) shows that trade reduces employment in production and, consequently, increases the long-run research activity in the foreign country, which implies that trade increases growth. In terms of Fig. 1, trade creates an outward shift in the DD schedule, leading to a higher interest rate and faster growth in equilibrium.

The result can be interpreted as trade leading to specialization. The home country specializes in final production, while the foreign country diversifies between manufacturing and innovation.[16] This is efficient, since there are country-wide economies of scale in innovation. Although trade leads to zero innovation in the home country, markets are integrated: final good producers, in both countries, can use the same varieties of intermediates and all consumers in the world can invest in the innovative firms of the foreign economy. Therefore, the location of innovation and firms has no impact on the relative welfare of the two countries.

Consider now the case when trade induces cross-country flows of ideas, i.e., if the knowledge spillover is determined, after trade, by the world stock of ideas contained in the union of A and A^*. When free trade is allowed, the accumulation of knowledge in each country is given by

$$\dot{A} = \delta L_x\big(A + A^*\big) \quad \text{and} \quad \dot{A}^* = \delta L_x^*\big(A + A^*\big).$$

Even if trade did not affect the allocation of the workforce between production and research, the rate of growth of technology would increase. But there is an additional effect; the larger knowledge spillover increases labor productivity in research, inducing an increase of employment in research. Formally, the total effect is equivalent to an increase in parameter δ. In terms of Figure 1, trade in goods plus flow of ideas imply an upward shift of the DD locus for both countries. Hence, trade attains the same effect as economic integration (increasing δ is equivalent to increasing L). This result is robust to asymmetric initial conditions.

3.2. Innovation, imitation and product cycles

The model just presented may be appropriate for describing trade integration between similar countries, but it misses important features of North–South trade. In a seminal article, Vernon (1966) argued that new products are first introduced in rich countries (the North), where R&D capabilities are high and the proximity to large and rich markets facilitates innovation. After some time, when a product reaches a stage of maturity and manufacturing methods become standardized, the good can easily be imitated and then, the bulk of production moves to less developed countries (the South), to take advantage of low wages. The expanding variety model provides a natural framework for studying the introduction of new goods and their subsequent imitation (product cycle trade).[17]

[16] Home-country patent holders will still produce intermediates, but as compared to the world's stock of intermediates, they will be of measure zero.

[17] Quality ladder models of innovation have been used to study product-cycles by, among others, Grossman and Helpman (1991a, 1991b), Segerstrom et al. (1990) and Dinopoulos and Segerstrom (2003).

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 18
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

We have already discussed imitation within the context of a closed economy. Here, we extend the analysis to the case where a richer North innovates, while a poorer South only engages in imitation. The analysis yields new results that modify some of the previous conclusions on the effect of imitation on innovation. The key questions are, first, how the transfer of production to the South through imitation affects the incentives to innovate and, second, how it affects the income distribution between North and South.

Following Helpman (1993), consider a two-region model of innovation, imitation and trade. Assume that R&D, producing new goods, is performed in the North only and that costless imitation takes place in the South at a constant rate m.[18] The imitation rate can be interpreted as an inverse measure of protection of Intellectual Property Rights (IPRs). Once a good is copied in the South, it is produced by competitive firms. Therefore, at every point in time, there is a range A_t^N of goods produced by monopolists in the North and a range A_t^S of goods that have been copied and are produced in the South by competitive firms. Given that the rate of introduction of new good is $\gamma = \dot{A}_t/A_t$, where $A_t = A_t^N + A_t^S$, and that monopolized goods are copied at the instantaneous rate m, $\dot{A}_t^S = mA_t^N$, it follows that a steady-state where the ratio A_t^N/A_t^S is constant must satisfy:

$$\frac{A^N}{A} = \frac{\gamma}{\gamma + m} \quad \text{and} \quad \frac{A^S}{A} = \frac{m}{\gamma + m}. \tag{25}$$

We use the "labor-for-intermediates" version of the growth model, so that the price of a single variety depends on the prevailing wage rate in the country where it is manufactured. This is an important feature of product cycle models, allowing the North to benefit from low production costs in the South for imitated goods. Therefore, we define the aggregate production function as in (18):

$$Y_t = \int_0^{A_t} {x_i}^{\alpha}\,\mathrm{d}i, \tag{26}$$

where A_t is the (growing) range of available products x_i and $\varepsilon = 1/(1-\alpha)$ is the elasticity of substitution between any two varieties. Intermediates are manufactured with $1/A_t$ units of labor per unit of output in both regions. Northern firms charge a monopoly price, as long as their products have not been imitated, equal to a constant markup $1/\alpha$ over the production cost, given by the wage rate. On the contrary, Southern firms produce imitated goods that have become competitive and sell them at a price equal to the marginal cost. To summarize:

$$p_t^N = \frac{w_t^N}{\alpha A_t} \quad \text{and} \quad p_t^S = \frac{w_t^S}{A_t}, \tag{27}$$

where p_t^N and p_t^S are the prices of any variety of intermediates produced in the North and South, respectively.

[18] The rate of imitation is made endogenous in Grossman and Helpman (1991b).

As in the benchmark model, innovation requires labor: the introduction of new products per unit of time $\dot{A}_t$ equals $\delta A_t L_x$, where L_x is the (Northern) labor input employed in R&D, δ is a productivity parameter and A_t captures an externality from past innovations. This implies that the growth rate of the economy is a linear function of the number of workers employed in R&D, $\gamma = L_x\delta$. As usual, profits generated by the monopoly over the sale of the new good are used to cover the cost of innovation. Since the profits per product are a fraction $(1-\alpha)$ of total revenue $p^N x$ and the labor market clears, $A_t^N x/A_t + \gamma/\delta = L^N$, profits can be written as:

$$\pi^N = \frac{1-\alpha}{\alpha}\frac{w_t^N}{A_t^N}\left(L^N - \frac{\gamma}{\delta}\right). \tag{28}$$

Arbitrage in asset markets implies that $(r+m)V^N = \pi^N + \dot{V}^N$, where V^N is the PDV of a new good and the effective interest rate is adjusted by the imitation risk. Along a BG path, $\dot{V}^N = 0$ and free entry ensures that the value of an innovation equals its cost, $w_t^N/\delta A_t$. Combining these considerations with (25) and (28) yields:

$$\frac{1-\alpha}{\alpha}\big(\delta L^N - \gamma\big)\frac{\gamma+m}{\gamma} = r+m. \tag{29}$$

Together with the Euler equation for consumption growth, (29) provides an implicit solution for the long-run growth rate of innovation. Note that the left-hand side is the profit rate (i.e., instantaneous profits over the value of the innovation) and the right-hand side represents the effective cost of capital, inclusive of the imitation risk.

To see the effect of a tightening of IPRs (a reduction of m), consider how an infinitesimal change in m affects the two sides of (29). Taking a log linear approximation, the impact of m on the profit rate is $1/(\gamma+m)$, whereas the effect on the cost of capital is $1/(r+m)$. In the case of log preferences, studied by Helpman (1993), $r > \gamma$. Hence, a reduction of m has a larger impact on the profit rate than on the effective cost of capital, thereby reducing the profitability of innovation and growth. What is the effect on the fraction of goods produced in the North? Rewriting (29) with the help of (25) as:

$$\frac{A^N}{A} = \frac{1-\alpha}{\alpha}\big(\delta L^N - \gamma\big)\frac{1}{r+m}, \tag{30}$$

it becomes apparent that a reduction of m increases the share of goods manufactured in the North, both through its direct effect and by reducing γ and r.

To understand these results, note that stronger IPRs have two opposite effects. First, a lower imitation rate prolongs the expected duration of the monopoly on a new product developed in the North, thereby increasing the returns to innovation. Second, since firms produce for a longer time in the North, it rises the demand for Northern labor, w^N, and hence, the cost of innovation. For the specification with log utility, the latter effect dominates and innovation declines. More generally, the link between the rate of imitation and innovation can go either way [as in Grossman and Helpman (1991a)]. However, the important result here is that tighter IPRs does not necessarily stimulate innovation in the long run.

The effect of IPRs on the North–South wage ratio can be found using (27), together with the relative demand for intermediates:[19]

$$\frac{w^N}{w^S} = \alpha\left[\frac{A^N}{A^S}\frac{L_S}{L_N - \gamma/\delta}\right]^{1/\varepsilon}. \tag{31}$$

Given that a decline in the imitation rate m raises $(A^N/A^S)/(L_N - \gamma/\delta)$ (see Equation (30)), a tightening of IPRs raises the relative wage of the North. Helpman (1993) computes welfare changes in the North and in the South (including transitional dynamics) after a change in the imitation rate m, and concludes that the South is unambiguously hurt by a decline in imitation. Moreover, if the imitation rate is not too high, the North can also be worse-off.

More recent papers on product cycles, incorporating the notion that stronger IPRs make relocation of production to the South a more attractive option, have come to different conclusions. For example, by assuming that Northern multinationals can produce in the South and that Southern firms can only imitate after production has been transferred to their country, Lai (1998) shows that stronger IPRs increase the rate of product innovation and the relative wage of the South. Similarly, Yang and Maskus (2001) find that if Northern firms can license their technology to Southern producers, being subject to an imitation risk, stronger IPRs reduce the cost of licensing, free resources for R&D and foster growth, with ambiguous effects on relative wages. Finally, the literature on appropriate technology [e.g., Diwan and Rodrik (1991), Acemoglu and Zilibotti (2001), Gancia (2003)] has shown that, when the North and the South have different technological needs, the South has an incentive to protect IPRs in order to attract innovations more suited to their technological needs. Some of these results are discussed in the next sections.

4. Directed technical change

So far, technical progress has been modeled as an increase in total factor productivity (A) that is neutral towards different factors and sectors. For many applications, however, this assumption is not realistic. For example, there is evidence that technical progress has been skill-biased during the last century and that this bias accelerated during the 1980s. Similarly, the fact that the output shares of labor and capital have been roughly constant in the US while the capital–labor ratio has been steadily increasing

[19] Relative demand for intermediates is:

$$\frac{p^N}{p^S} = \left(\frac{x^N}{x^S}\right)^{\alpha-1}.$$

Using $x^N = AL_y^N/A^N$, $x^S = AL_y^S/A^S$ and the pricing formula (27) yields the expression in the text.

suggests that technical change has mainly been labor-augmenting.[20] Further, industry studies show R&D intensity to vary substantially across sectors. In order to build a theory for the direction of technical change, a first step is to introduce more sectors into the model. Then, studying the economic incentives to develop technologies complementing a specific factor or sector can help understand what determines the shape of technology.

An important contribution of this new theory will be to shed light on the determinants of wage inequality [Acemoglu (1998, 2003a)]. Another application studies under which circumstances technologies developed by profit-motivated firms are appropriate for the economic conditions of the countries where they are used. The analysis will demonstrate that, since IPRs are weakly protected in developing countries, new technologies tend to be designed for the markets and needs of advanced countries. As a result, these technologies yield a low level of productivity when adopted by developing countries [Acemoglu and Zilibotti (2001)]. Trade can reinforce this problem and create interesting general equilibrium effects.

Although most of the results discussed in this section can be derived using models of vertical innovation, the expanding variety approach has proved to be particularly suited for addressing these issues because of its analytical tractability and simple dynamics. For instance, creative destruction, a fundamental feature of quality-ladder models, is not a crucial element for the problems at hand, and abstracting from it substantially simplifies the analysis.

4.1. Factor-biased innovation and wage inequality

Directed technical change was formalized by Acemoglu (1998), and then integrated by Acemoglu and Zilibotti (2001) into a model of growth with expanding variety, to explain the degree of skill-complementarity of technology.[21] In this section, we discuss the expanding variety version [following the synthesis of Acemoglu (2002)] by extending the "lab-equipment" model to two sectors employing skilled and unskilled labor, respectively. Consider the following aggregate production function:

$$Y = \left[Y_L^{(\varepsilon-1)/\varepsilon} + Y_H^{(\varepsilon-1)/\varepsilon}\right]^{\varepsilon/(\varepsilon-1)}, \tag{32}$$

where Y_L and Y_H are goods produced with unskilled labor, L, and skilled labor, H, respectively. Y represents aggregate output, used for both consumption and investment, as a combination of the two goods produced in the economy, with an elasticity of substitution equal to ε. Maximizing Y under a resource constraint gives constant elasticity demand functions, implying a negative relationship between relative prices and relative

[20] Unless the production function is Cobb–Douglas, in which case the direction of technical progress is irrelevant. Empirical estimates suggest that the elasticity of substitution between labor and capital is likely to be less than one. See Hamermesh (1993) for a survey of early estimates and Krusell et al. (2000) and Antras (2004) for more recent contributions.

[21] Important antecedents are Kennedy (1964), Samuelson (1965) and Atkinson and Stiglitz (1969).

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 22
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

quantities:

$$\frac{P_H}{P_L} = \left[\frac{Y_L}{Y_H}\right]^{1/\varepsilon}, \tag{33}$$

where P_L and P_H are the prices of Y_L and Y_H, respectively. Aggregate output is chosen as the numeraire, hence:

$$\left(P_L^{1-\varepsilon} + P_H^{1-\varepsilon}\right)^{1/(1-\varepsilon)} = 1. \tag{34}$$

The distinctive feature of this model is that the two goods are now produced using different technologies:

$$\begin{aligned} Y_L &= L^{1-\alpha} \int_0^{A_L} {x_{L,j}}^{\alpha}\, \mathrm{d}j, \\ Y_H &= H^{1-\alpha} \int_0^{A_H} {x_{H,j}}^{\alpha}\, \mathrm{d}j, \end{aligned} \tag{35}$$

where $x_{L,j}$, $j \in [0, A_L]$, are intermediate goods complementing unskilled labor, whereas $x_{H,j}$, $j \in [0, A_H]$ complement skilled labor. This assumption captures the fact that different factors usually operate with different technologies and that a new technology may benefit one factor more than others.[22] For example, it has been argued that computers boosted the productivity of skilled more than that of unskilled labor, whereas the opposite occurred after the introduction of the assembly line. As before, technical progress takes the form of an increase in the number of intermediate goods, $[A_L, A_H]$, but now an innovator must decide which technology to expand. The profitability of the two sectors pins down, endogenously, the direction of technical change. In a steady-state equilibrium, there is a constant ratio of the number of intermediates used by each factor, A_H/A_L, and this can be interpreted as the extent of the "endogenous skill-bias" of the technology.

The analysis follows the same steps as in model with a single factor. Final good producers take the price of their output (P_L, P_H), the price of intermediates $(p_{L,j}, p_{H,j})$ and wages (w_L, w_H) as given. Consider a variety j used in the production of Y_L. Profit maximization gives the following isoelastic demand:

$$x_{L,j} = \left[\frac{\alpha P_L}{p_{L,j}}\right]^{1/(1-\alpha)} L, \tag{36}$$

and an equivalent expression for $x_{H,j}$.

The intermediate good sector is monopolistic, with each producer owning the patent for a single variety. The cost of producing one unit of any intermediate good is one

[22] The analysis can be generalized to specifications where, in the spirit of Heckscher–Ohlin models, each sector uses all productive factors, but factor intensities differ across sectors. The model can also be generalized to more than two factors and sectors.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 23
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

unit of the numeraire. The symmetric structure of demand and technology implies that all monopolists set the same price, $p_{Lj} = p_L$. In particular, given the isoelastic demand, they set $p_L = 1/\alpha$ and sell the quantity $x_{L,j} = (\alpha^2 P_L)^{1/(1-\alpha)} L$. The profit flow accruing to intermediate producers can therefore be expressed as

$$\pi_L = (1-\alpha)\alpha^{(1+\alpha)/(1-\alpha)} (P_L)^{1/(1-\alpha)} L. \tag{37}$$

Similar conclusions are reached for varieties used in the production of Y_H, leading to

$$\pi_H = (1-\alpha)\alpha^{(1+\alpha)/(1-\alpha)} (P_H)^{1/(1-\alpha)} H. \tag{38}$$

From (37)–(38), it immediately follows that the relative profitability in the two sectors is given by

$$\frac{\pi_H}{\pi_L} = \left(\frac{P_H}{P_L}\right)^{1/(1-\alpha)} \frac{H}{L}, \tag{39}$$

which, since profits are used to finance innovation, is also the relative profitability of R&D directed to the two sectors. The first term in (39) represents the "price effect": there is a greater incentive to invent technologies producing more expensive goods.[23] The second term is the "market size" effect: the incentive to develop a new technology is proportional to the number of workers that will be using it.[24]

Next, using the price of intermediates in (36) and (35) gives final output in each sector:

$$\begin{aligned} Y_L &= \alpha^{2\alpha/(1-\alpha)} P_L^{\alpha/(1-\alpha)} A_L L, \\ Y_H &= \alpha^{2\alpha/(1-\alpha)} P_H^{\alpha/(1-\alpha)} A_H H. \end{aligned} \tag{40}$$

Note the similarity with (16). As in the benchmark model, output – in each sector – is a linear function of technology and labor. But sectoral output now also depends on sectoral prices, P_L and P_H, since a higher price of output increases the value of productivity of intermediates, but not their costs, and therefore encourages firms to use more of them, thereby raising labor productivity. Note that this is not the case in the one-sector model since there, the price of output is proportional to the price of intermediates.

We can now solve for prices and wages as functions of the state of technology and endowments. Using (40) into (33) and noting that the wage bill is a constant fraction of sectoral output, yields:

$$\frac{P_H}{P_L} = \left[\frac{A_H}{A_L}\frac{H}{L}\right]^{-(1-\alpha)/\sigma}, \tag{41}$$

$$\frac{w_H}{w_L} = \left[\frac{A_H}{A_L}\right]^{1-1/\sigma} \left[\frac{H}{L}\right]^{-1/\sigma}, \tag{42}$$

[23] The price effect, restated in terms of factor prices, was emphasized by Hicks (1932) and Habakkuk (1962).

[24] Market size, although in the context of industry- and firm-level innovation, was emphasized as a determinant of technical progress by Griliches and Schmookler (1963), Schmookler (1966) and Schumpeter (1950).

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 24
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

where $\sigma \equiv 1 + (1-\alpha)(\varepsilon - 1)$ is, by definition, the elasticity of substitution between H and L.[25] Note that the skill premium, w_H/w_L, is decreasing in the relative supply of skilled labor (H/L) and increasing in the skill-bias (A_H/A_L).

The final step is to find the equilibrium for technology. We assume, as in the lab-equipment model of Section 2.2, that the development of a new intermediate good to require a fixed cost of μ units of the numeraire. Free entry and an arbitrage condition require the value V_z of an innovation directed to factor $Z \in \{L, H\}$ to equal its cost. Since the value of an innovation is the PDV of the infinite stream of profits it generates, an equilibrium with a positive rate of innovation in both types of intermediates such that the ratio A_H/A_L remains constant, i.e., a BG path where P_H/P_L, w_H/w_L and π_H/π_L are also constant, requires profit equalization in the two sectors, $\pi_H = \pi_L = \pi$. Imposing this restriction yields the equilibrium skill-bias of technology,

$$\frac{A_H}{A_L} = \left[\frac{H}{L}\right]^{\sigma-1}. \tag{43}$$

Equation (43) shows that, as long as workers of different skill levels are gross substitutes ($\sigma > 1$), an increase in the supply of one factor will induce more innovation directed to that specific factor. This is the case because, with $\sigma > 1$, the market size effect dominates the price effect, and technology is biased towards the abundant factor. The opposite is true if $\sigma < 1$. As usual, the growth rate of the economy can be found from the free-entry condition $\pi_Z/r = \mu$, $Z \in \{L, H\}$. Using (34), (41) and (43) to substitute for prices and the interest rate from the Euler equation, yields:

$$\gamma = \frac{1}{\theta}\left[\frac{(1-\alpha)\alpha^{(1+\alpha)/(1-\alpha)}}{\mu}\left(L^{\sigma-1} + H^{\sigma-1}\right)^{1/(\sigma-1)} - \rho\right].$$

If we only had one factor (e.g., $H = 0$), the growth rate would reduce to that of the benchmark model.

Directed technical change has interesting implications on factor prices. Using (43), the skill-premium becomes:

$$\frac{w_H}{w_L} = \left[\frac{H}{L}\right]^{\sigma-2}. \tag{44}$$

Equation (44) shows that the slope of the labor demand curve, i.e., the relationship between relative wages and relative labor supply, can be either positive or negative and is the result of two opposite forces. On the one hand, a large supply of one factor depresses the price of its product while, on the other hand, it induces a technology bias in its favor, thereby raising its productivity. A high substitutability between H and L implies a weak price effect of an increase in relative supply, which makes a positive relationship more likely. In particular, if $\sigma > 2$, the market size effect is sufficiently strong to not

[25] This is the short-run elasticity of substitution between L and H, for a given technology A_L and A_H.

only dominate the price effect on technical change (see Equation (43)), but also the substitution effect between skilled and unskilled workers at a given technology.

This result can help rationalize several facts. First, it suggests that technical change has been skill biased during the past 60 years, because of the steady growth in the supply of skilled labor. Second, the case $\sigma > 2$ offers an explanation for the fall and rise in the US skill premium during the 1970s and 1980s. In the 1970s, there was a large increase in the supply of skilled labor (H/L). Assuming this shock to be unexpected, the model predicts an initial fall in the skill premium (recall that A_H/A_L is a state variable that does not immediately adjust), followed by its rise due to the induced skill biased technical change, a pattern broadly consistent with the evidence.

In Acemoglu (2003b), this set-up is used to study the direction of technical progress when the two factors of production are capital and labor. Beyond the change of notation, the resulting model has an important qualitative difference, as capital can be accumulated. The main finding is that, when both capital and labor augmenting innovations are allowed, a balanced growth path still exists and features labor-augmenting technical progress only. The intuition is that, while there are two ways of increasing the production of capital-intensive goods (capital-augmenting technical change and accumulation), there is only one way of increasing the production of labor-intensive goods (labor-augmenting technical progress). Therefore, in the presence of capital accumulation, technical progress must be more labor-augmenting than capital-augmenting. Further, Acemoglu shows that, if capital and labor are gross complements (i.e., the elasticity of substitution between the two is less than one), which seems to be the empirically relevant case [see, for example, Antras (2004)], the economy converges to the balanced growth path.

Finally, the theory of directed technical change can be used to study which industries attract more innovation and why R&D intensity differs across sectors. In this exercise, following a modified version of Klenow (1996), we abstract from factor endowments as determinants of technology, by assuming there to be a single primary input, which we call labor. Instead, other characteristics can make one sector more profitable than others. Major explanations put forward in the literature on innovation are industry differences in technological opportunities, market size and appropriability of rents, all factors that can easily be embedded in the basic model with two sectors. In particular, to capture the market size hypothesis, we introduce a parameter η defining the relative importance of industry i in aggregate consumption:

$$Y = \left[\eta Y_i^{(\varepsilon-1)/\varepsilon} + (1-\eta) Y_j^{(\varepsilon-1)/\varepsilon}\right]^{\varepsilon/(\varepsilon-1)}.$$

Differences in technological opportunities can be incorporated by allowing the cost of an innovation, μ_i, to vary across sectors. Finally, we assume that an inventor in industry i can only extract a fraction λ_i of the profits generated by his innovation. The previous analysis carries over almost unchanged, with the main difference that we now need to solve for the allocation of labor across industries. This can be done requiring all

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 26
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

industries to pay the same wage, i.e., setting (42) equal to one:

$$\frac{L_i}{L_j} = \left(\frac{A_i}{A_j}\right)^{\sigma-1}.$$

Solving the new arbitrage condition stating that innovation for the two industries should be equally profitable in BG, $\lambda_i \pi_i/\mu_i = \lambda_j \pi_j/\mu_j$, yields the relative industry-bias of technology:

$$\frac{A_i}{A_j} = \left(\frac{\lambda_i}{\lambda_j}\frac{\mu_j}{\mu_i}\right)^{1/(2-\sigma)} \left(\frac{\eta}{1-\eta}\right)^{\varepsilon/(2-\sigma)}.$$

As expected, industries with a larger market size, better technological opportunities and higher appropriability attract more innovations.[26] Empirical estimates surveyed by Cohen and Levin (1989) suggest that about one half of the industry differences in research intensity can be attributed to the available measures of these three factors.

4.2. Appropriate technology and development

Directed technical change has interesting implications for the analysis of some development issues. Acemoglu and Zilibotti (2001) show that technologies resulting from directed technical change are optimal for the economic conditions of the markets where they are sold. They analyze the implications of this finding in a two-country world where technological innovation takes place in the North, and the South does not enforce (or imperfectly enforce) IPRs. In this environment, innovators in the North can only extract rents from selling technologies (embodied in new varieties of intermediate goods) in the Northern market, since new technologies can be copied and locally produced in the South. Thus, innovation does not respond to the factor endowment of the South: the equilibrium skill-bias of technical change (see Equation (43) in the previous section) is determined by the factor endowment of the North only. In this sense, technological development tends to be "inappropriate" for the South: there is too much investment in inventing new technologies augmenting the productivity of skilled workers, and too little in inventing new technologies augmenting the productivity of unskilled workers. Such excessive skill-bias prevents the South from fully profiting from technological improvements. The theory can explain North–South productivity differences, even when the technology is identical and there are no significant barriers to technology adoption.[27]

[26] This is true as long as $\sigma < 2$. This restriction is required to have balanced growth. If violated, e.g., if goods are highly substitutable, it would be profitable to direct innovation to one sector only.

[27] Evidence on cross-country TFP differences is provided by, among others, Klenow and Rodriguez-Clare (1997), Hall and Jones (1999), Caselli, Esquivel and Lefort (1996) and Prescott (1998). The view that technological differences arise from barriers to technology adoption is expressed by, among others, Parente and Prescott (1994) and Prescott (1998).

We start by studying the set of advanced countries, called North. A continuum of measure one of final goods is produced by competitive firms. Final goods, indexed by $i \in [0,1]$, are aggregated to give a composite output, $Y = \exp(\int_0^1 \log y_i \, \mathrm{d}i)$, which is the numeraire. There are two differences with respect to the model of the previous section: first, there is a continuum of sectors, not just two, and second, the elasticity of substitution between sectors is unity.[28] Each good i can be produced with both skilled and unskilled labor using two sets of intermediate goods: intermediates $[0, A_L]$ used by unskilled workers only and intermediates $[0, A_H]$ used by skilled workers only. Therefore, despite the continuum of sectors, there are only two types of technologies, as in the basic model of directed technical change. The production function takes the following form:

$$y_i = \big[(1-i)l_i\big]^{1-\alpha} \int_0^{A_L} x_{L,v,i}^{\alpha} \, \mathrm{d}v + [ih_i]^{1-\alpha} \int_0^{A_H} x_{H,v,i}^{\alpha} \, \mathrm{d}v, \tag{45}$$

where l_i and h_i are the quantities of unskilled and skilled labor employed in sector i, respectively, and $x_{z,v,i}$ is the quantity of intermediate good of type v used in sector i together with the labor of skill level $z = L, H$. Note that sectors differ in labor-augmenting productivity parameters, $(1-i)$ for the unskilled technology and i for the skilled technology, so that unskilled labor has a comparative advantage in sectors with a low index. Producers of good i take the price of their product, P_i, the price of intermediates ($p_{L,v}$, $p_{H,v}$) and wages (w_L, w_H) as given. Profit maximization gives the following demands for intermediates:

$$x_{L,v,i} = (1-i)l_i[\alpha P_i / p_{L,v}]^{1/(1-\alpha)} \quad \text{and} \quad x_{H,v,i} = ih_i[\alpha P_i / p_{H,v}]^{1/(1-\alpha)}. \tag{46}$$

The intermediate good sector is monopolistic. Each producer holds the patent for a single type of intermediate good v, and sells its output to firms in the final good sectors. The cost of producing one unit of any intermediate is conveniently normalized to α^2 units of the numeraire. Profit maximization by monopolists implies that prices are a constant markup over marginal costs, $p = \alpha$. Using the price of intermediates together with (46) and (45) gives the final output of sector i as a linear function of the number of intermediate goods and labor:

$$y_i = P_i^{\alpha/(1-\alpha)}\big[A_L(1-i)l_i + A_H i h_i\big]. \tag{47}$$

From (47), it is easily seen that all sectors whose index i is below a threshold level J will use the unskilled technology only and the remaining sectors will employ the skilled technology only. This happens because of the comparative advantage of unskilled workers in low index sectors and the linearity of the production function (there is no incentive to combine the two technologies and, for a given i, one always dominates the other).

[28] The composite output Y can be interpreted as a symmetric Cobb–Douglas over the measure of final goods $i \in [0,1]$.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 28
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

The total profits earned by monopolists are:

$$\pi_{L,v} = (1-\alpha)\alpha \int_0^1 P_i^{1/(1-\alpha)}(1-i)l_i \,\mathrm{d}i \quad \text{and}$$
$$\pi_{H,v} = (1-\alpha)\alpha \int_0^1 P_i^{1/(1-\alpha)} i h_i \,\mathrm{d}i. \tag{48}$$

Note that, by symmetry, $\pi_{L,v} = \pi_{L,j}$ and $\pi_{H,v} = \pi_{H,j}$. Given the Cobb–Douglas specification in (45), the wage bill in each sector is a fraction $(1-\alpha)$ of sectoral output. Therefore, Equation (47) can be used to find wages:[29]

$$w_L = (1-\alpha)P_i^{1/(1-\alpha)}A_L(1-i) \quad \text{and} \quad w_H = (1-\alpha)P_i^{1/(1-\alpha)}A_H i. \tag{49}$$

Defining $P_L \equiv P_0$, $P_H \equiv P_1$ and dividing equations in (49) by their counterparts in sectors 0 and 1, respectively, it is possible to derive the following pattern of prices: for $i \leqslant J$, $P_i = P_L(1-i)^{-(1-\alpha)}$ and for $i \geqslant J$, $P_i = P_H i^{-(1-\alpha)}$. Intuitively, the price of a good produced with skilled (unskilled) labor is decreasing in the sectoral productivity of skilled (unskilled) workers. Next, note that to maximize Y, expenditures across goods must be equalized, i.e., $P_i y_i = P_H y_1 = P_L y_0$ (as for a symmetric Cobb–Douglas). This observation, plus the given pattern of prices and full employment, imply that labor is evenly distributed among sectors: $l_i = L/J$, $h_i = H/(1-J)$, as prices and sectoral productivity compensate each other. Finally, in sector $i = J$, it must be the case that both technologies are equally profitable or $P_L(1-J)^{-(1-\alpha)} = P_H J^{-(1-\alpha)}$; this condition, using $P_H y_1 = P_L y_0$ and (47), yields:

$$\frac{J}{1-J} = \left(\frac{P_H}{P_L}\right)^{1/(1-\alpha)} = \left(\frac{A_H}{A_L}\frac{H}{L}\right)^{-1/2}. \tag{50}$$

The higher the relative endowment of skill (H/L) and the skill-bias of technology (A_H/A_L), the larger the fraction of sectors using the skill-intensive technology $(1-J)$. Finally, integrating $P_i y_i$ over [0, 1], using (47), (50) and the fact that the consumption aggregate is the numeraire (i.e., $\exp[\int_0^1 \ln P_i \,\mathrm{d}i] = 1$) gives a simple representation for aggregate output:

$$Y = \exp(-1)\left[(A_L L)^{1/2} + (A_H H)^{1/2}\right]^2, \tag{51}$$

which is a CES function of technology and endowments, with an elasticity of substitution between factors equal to two.

So far, the analysis defines an equilibrium for a given technology. Next, we need to study innovation and characterize the equilibrium skill-bias of technology, (A_H/A_L). As before, technical progress takes the form of an increase in A_L and A_H and is the

[29] In Acemoglu and Zilibotti (2001), there is an additional parameter ($Z > 1$), which is here omitted for simplicity, which augments the productivity of skilled workers, ensuring that the skill premium is positive in equilibrium.

result of directed R&D investment. The cost of an innovation (of any type) is equal to μ units of the numeraire, and R&D is profitable as long as the PDV of the infinite flow of profits that a producer of a new intermediate expects to earn covers the fixed cost of innovation. Finally, free entry ensures that there are no additional profits. Using the price pattern, instantaneous profits can be simplified as:

$$\pi_H = \alpha(1-\alpha)P_H^{1/(1-\alpha)}H. \tag{52}$$

A parallel expression gives π_L. Balanced growth requires $\pi_L = \pi_H$; in this case, A_H and A_L grow at the same rate, the ratio A_H/A_L is constant as are J, P_L and P_H. Imposing $\pi_L = \pi_H$ in (52) and using (50) yields:

$$\frac{A_H}{A_L} = \frac{1-J}{J} = \frac{H}{L}. \tag{53}$$

Note that the equilibrium skill-bias is identical to that of (43) in the special case when $\sigma = 2$. Further, (53) shows that the higher is the skill endowment of a country, the larger is the range of sectors using the skilled technology. This is a complete characterization of the equilibrium for fully integrated economies developing and selling technologies in their markets with full protection of IPRs and can be interpreted as a description of the collection of rich countries, here called the North.

Consider now Southern economies, where skilled labor is assumed to be relatively more scarce: $H^S/L^S < H^N/L^N$. Assume that intellectual property rights are not enforced in the South and that there is no North–South trade. It follows that intermediate producers located in the North cannot sell their goods or copyrights to firms located in the South, so that the relevant market for technologies is the Northern market only. Nonetheless, Southern producers can copy Northern innovations at a small but positive cost. As a consequence, no two firms in the South find it profitable to copy the same innovation and all intermediates introduced in the North are immediately copied (provided that the imitation cost is sufficiently small) and sold to Southern producers by a local monopolist. Under these assumptions, firms in the South take the technologies developed originating in the North as given and do not invest in innovation.[30] This means that both the North and the South use the same technologies, but $A_H/A_L = H^N/L^N$, i.e., the skill-bias is determined by the factor endowment of the North, since this is the only market for new technologies. Except for this, the other equilibrium conditions also apply to the South after substituting the new endowments, H^S and L^S.

We are now ready to answer the following questions: are technologies appropriate for the skill endowment of the countries where they are developed? What happens to aggregate productivity if they are used in a different economic environment?

[30] Imitation can be explicitly modelled as an activity similar to innovation, but less costly. Assuming the cost of an innovation of type z to decrease with the distance from the relevant technology frontier A_Z^N, as in Barro and Sala-i-Martin (1997), would yield very similar results.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 30
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

Simple differentiation on (51) establishes that Y is maximized for $A_H/A_L = H/L$. This is exactly condition (53), showing that the equilibrium skill-bias is optimally chosen for the Northern skill composition. On the contrary, since factor abundance in the South does not affect the direction of technical change, new technologies developed in the North are inappropriate for the needs of the South. As a consequence, output per capita, $Y/(L + H)$ is greater in the North than in the South. The reason for these productivity differences is a technology–skill mismatch. To understand why, note that, from Equation (50), $J^S > J^N$. Rewriting (53) as $A_H J^N = A_L(1 - J^N)$ and inspecting Equation (47) reveals that unskilled workers are employed in the North up to sector J^N, where they become as productive as skilled workers. This basic efficiency condition is violated in the South, where $A_H J^S > A_L(1 - J^S)$. Because of its smaller skill endowment, the South is using low-skill workers in some sectors where high-skill workers would be more productive.

This result can help understand the existence of substantial differences in TFP across countries, even when the technology is common. In particular, Acemoglu and Zilibotti (2001) compare the predictive power of their model in explaining cross-country output differences with that of a comparable neoclassical model, where all countries have access to the same technologies and output is Cobb–Douglas in labor, human and physical capital. Their computations suggest that the proposed mechanism can account for one-third to one half of the total factor productivity gap between the United States and developing countries. Predictions on the pattern of North–South, cross-industry, productivity differences are also tested. Since the South uses the same technology $[A_L, A_H]$ as the rest of the world, but it has a higher relative price for skill-intensive goods, it follows that the value of productivity in LDCs relative to that of the North should be higher in skill-intensive sectors. The empirical analysis supports this prediction.

The view that countries adopt different technologies out of a world "menu", and that the choice of the appropriate technology depends on factor endowments, particularly on the average skill of the labor force, finds support in the analysis of Caselli and Coleman (2000). However, these authors also find that many poor countries choose technologies inside the world technology frontier, thereby suggesting that barriers to technology adoption may also be important to explain the low total factor productivity of these countries.

4.3. Trade, inequality and appropriate technology

We have seen that directed technical change can help understand inequality, both within and between countries. Several authors have stressed that international trade is another important determinant of income distribution. For example, Wood (1994) argues that the higher competition with imports from LDCs may be responsible for the deterioration in relative wages of low-skill workers in the US in the past decades. Further, there is a widespread concern that globalization may be accompanied by a widening of income differences between rich and poor countries. Although the analysis of these issues goes beyond the scope of this paper, we want to argue that R&D-driven endoge-

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 31
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

nous growth models can fruitfully be used to understand some of the links between trade and inequality. In particular, we now show that trade with LDCs can have a profound impact on income distribution, beyond what is suggested by static trade theory, through its effect on the direction of technical change. By changing the relative prices and the location of production, international trade can change the incentives for developing innovations targeted at specific factors or sectors, systematically benefiting certain groups or countries more than others. A key assumption in deriving these results is that, as in the previous paragraph, LDCs do not provide an adequate protection of IPRs.

First, consider the effect of trade in the benchmark model of directed technical change. The analysis follows Acemoglu (2002, 2003a). Recall that the profitability of an innovation depends on its market size and the price of the goods it produces, as in Equation (39). What happens to technology if we allow free trade in Y_L and Y_H between a skill-abundant North and a skill-scarce South? The market size for innovations does not change, because inventors continue to sell their machines in the North only. But trade, at first, will increases the relative price of skill-intensive goods in the North. To see this, note that trade generates a single world market with a relative price depending on the world supply of goods. Since skills are scarcer in the world economy than in the North alone, trade will increase the relative price of skill-intensive goods in the North (the opposite will happen in the South). In particular, world prices are now given by Equation (41) using world endowments:

$$\frac{P_H}{P_L} = \left(\frac{A_H}{A_L}\frac{H^W}{L^W}\right)^{-(1-\alpha)/\sigma}. \tag{54}$$

This change in prices, for a given technology, makes skill-complement innovations more profitable and accelerates the creation of skill-complementary machines. Since, along the BG path, both types of innovations must be equally profitable and hence $\pi_H = \pi_L$, Equation (39) shows that this process continues until the relative price of goods has returned to the pre-trade level in the North. Substituting Equation (54) into (39) and imposing $\pi_H = \pi_L$, yields the new equilibrium skill bias of technology:

$$\frac{A_H}{A_L} = \frac{L^W}{H^W}\left[\frac{H^N}{L^N}\right]^{\sigma}. \tag{55}$$

Given that $H^N/L^N > H^W/L^W$, the new technology is more skill-biased and skilled workers in the North earn higher wages. The effect on the skill premium can be seen by substituting (55) into (42):

$$\frac{w_H}{w_L} = \left[\frac{H^N}{L^N}\frac{L^W}{H^W}\right]\left[\frac{H^N}{L^N}\right]^{\sigma-2}. \tag{56}$$

The effect of a move from autarky to free trade can be approximated by the elasticity of the skill premium to a change in L^W/H^W computed at $L^W/H^W = L^N/H^N$ (that is, starting from the pre-trade equilibrium). Equation (56) shows this elasticity to be unity.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 32
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

Thus, if, for example, L^W/H^W were 4% higher than L^N/H^N, the model would predict trade to raise the skill premium by the same 4%.[31]

Without technical change, instead, the reaction of the skill premium to a change in the perceived scarcity of factors due to trade depends on the degree of substitutability of skilled and unskilled workers. From Equation (42), the elasticity of the skill premium to a change in L/H would be $1/\sigma$, less than in the case of endogenous technology as long as $\sigma > 1$, i.e., when skilled and unskilled workers are gross substitutes. Therefore, with directed technical change and $\sigma > 1$, trade increases the skill premium in the North by more than would otherwise be the case: for example, if the elasticity of substitution is 2, the endogenous reaction of technical progress doubles the impact of trade on wage inequality.

Note that another direct channel through which trade can affect factor prices in models of endogenous technical change is by affecting the reward to innovation. If trade increases the reward to innovation (for example, through the scale effect) and the R&D sector is skill-intensive relative to the rest of the economy, trade will naturally spur wage inequality. This mechanism is studied by Dinopoulos and Segerstrom (1999) in a quality-ladder growth model with no scale effects.[32]

What are the implications of trade opening for cross-country income differences? We have seen that trade induces a higher skill bias in technology; given the result of Acemoglu and Zilibotti (2001) that the excessive skill-complementarity of Northern technologies is a cause of low productivity in Souther countries, it may seem natural to conclude that trade would then increase productivity differences. However, this conclusion would be premature. In the absence of any barriers, trade equalizes the price of goods; given that the production functions adopted so far rule out complete specialization, this immediately implies that factor prices and sectoral productivity are also equalized. This does not mean that trade equalizes income levels; because of their different skill-composition, the North and the South will still have differences in income per capita, but nothing general can be said.[33]

The fact that trade generates productivity convergence crucially depends on factor prices being equalized by trade. Since factor price equalization is a poor approximation of reality, it is worth exploring the implications of models with endogenous technologies when this property does not hold. A simple way of doing this is to add Ricardian productivity differences, so that trade opening leads to complete specialization. In this

[31] Borjas, Freeman and Katz (1997) show that 4% is a plausible estimate of the increase in the unskilled labor content of US trade with LDCs between 1980 and 1995. Therefore, this simple exercise may give a sense of how much of the roughly 20% increase in the US skill premium in the same period can be attributed to trade.

[32] Recently, other papers have suggested that trade between identical countries may as well increase skill premia through its effect on technology. See, for example, Epifani and Gancia (2002), Neary (2003) and Thoenig and Verdier (2003).

[33] A general result is that the endogenous response of technology makes trade less beneficial for LDCs than would otherwise be the case. This occurs because, after trade opening, the skill premium rises as a result of the induced skill-biased technical change. Given that the North is more skilled-labor abundant, it proportionally benefits more from a higher skill premium.

case, the endogenous response of technology to weak IPRs in LDCs becomes a force promoting productivity divergence.[34] Further, trade with countries providing weak protection for IPRs may have an adverse effect on the growth rate of the world economy. These results, shown by Gancia (2003), can be obtained by modifying Acemoglu and Zilibotti (2001) as follows. First, we allow the elasticity of substitution between final goods to be larger than one: $Y = [\int_0^1 y_i{}^{(\varepsilon-1)/\varepsilon}\,\mathrm{d}i]^{\varepsilon/(\varepsilon-1)}$, with $\varepsilon > 1$. Then, we assume that each good y_i can be produced by competitive firms both in the North and the South, using sector-specific intermediates and labor:

$$y_i = \left[(1-i)l_i^S\right]^{1-\alpha}\int_0^{A_i}\left(x_{i,v}^S\right)^{\alpha}\mathrm{d}v + \left[il_i^N\right]^{1-\alpha}\int_0^{A_i}\left(x_{i,v}^N\right)^{\alpha}\mathrm{d}v. \tag{57}$$

There are three important differences with respect to (45). First, $(1-i)$ and i now capture Ricardian productivity differences between the North and South, implying that the North is relatively more productive in high index sectors. Second, intermediate goods are sector specific, not factor specific (there is now a continuum $[0, 1]$ of technologies, not only two). Third, there is only one type of labor. Given that the endogenous component of technology (A_i) is still assumed to be common across countries, the sectoral North–South productivity ratio only depends on the Ricardian elements. The new implication is that countries specialize completely under free trade, as each good is only manufactured in the location where it can be produced at a lower cost.

The equilibrium can be represented by the intersection of two curves, as in Dornbusch et al. (1977). For any relative wage, the first curve gives the range $[0, J]$ of goods efficiently produced in the South: $\frac{J}{1-J} = \frac{w^N}{w^S}$. The second curve combines trade balance and a BG research arbitrage condition, requiring profits to be equalized across sectors and countries. To find this, the model assumes that the owner of a patent can only extract a fraction $\lambda < 1$ of the profits generated by its innovation in the South, so that λ can be interpreted as an index of the strength of international IPRs protection. The trade balance plus the research arbitrage condition turn out to be [see Gancia (2003)]:

$$\frac{w^N}{w^S} = \lambda^{-\widetilde{\sigma}}\left[\frac{L^S}{L^N}\frac{\int_J^1 (i)^{\widetilde{\sigma}/(1-\widetilde{\sigma})}\,\mathrm{d}i}{\int_0^J (1-i)^{\widetilde{\sigma}/(1-\widetilde{\sigma})}\,\mathrm{d}i}\right]^{1-\widetilde{\sigma}}, \tag{58}$$

with $\widetilde{\sigma} \equiv (1-\alpha)(\varepsilon-1) \in (0, 1)$.[35] As long as $\widetilde{\sigma} > 0$ (i.e., $\varepsilon > 1$), the wage gap is decreasing in the degree of protection of IPRs in the South, λ. The reason is that weaker protection of IPRs shifts innovations out of Southern sectors and increases the relative productivity of the North. From the condition $\frac{J}{1-J} = \frac{w^N}{w^S}$, it is easily seen that a weaker

[34] The idea that trade may magnify cross-country inequality was put forward by several economists. Some examples are Stiglitz (1970), Young (1991), Krugman and Venables (1995), Matsuyama (1996), Rodriguez-Clare (1996) and Ventura (1997).

[35] $\widetilde{\sigma} < 1$ guarantees balanced growth across sectors. $\widetilde{\sigma} > 0$, i.e., an elasticity of substitution between goods greater than one, rules out immiserizing growth.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 34
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

protection of IPRs in the South, by raising w^N/w^S, is accompanied by a reduction in sectors $[1-J]$ located in the North, because higher wages make the North less competitive. A second result emerges by calculating the growth rate of the world economy. In particular, Gancia (2003) shows the growth rate of the world economy to fall with λ and approach zero if λ is sufficiently low. The reason is that a lower λ shifts innovation towards Northern sectors and, at the same time, induces the relocation of more sectors to the South, where production costs become lower. This, in turn, implies that a wider range of goods becomes subject to weak IPRs and hence, to a low innovation incentive.

5. Complementarity in innovation

In the models described so far, innovation has no effect on the profitability of existing intermediate firms. This is a knife-edge property which descends from the specification of the final production technology, (3). In general, however, new technologies can substitute or complement existing technologies.

Innovation often causes technological obsolescence of previous technologies. Substitution is emphasized, in an extreme fashion, by Schumpeterian models such as Aghion and Howitt (1992). In such models, innovation provides "better of the same", i.e., more efficient versions of the pre-existing inputs. Growth is led by a process of creative destruction, whereby innovations do not only generate but also destroy rents over time. This has interesting implications for dynamics: the expectation of future innovations discourages current innovation, since today's innovators expect a short life of their rents due to rapid obsolescence. More generally, substitution causes a decline in the value of intermediate firms over time, at a speed depending on the rate of innovation in the economy.

There are instances, however, where new technologies complement rather than substitute old technologies. The market for a particular technology is often small at the moment of its first introduction. This limits the cash-flow of innovating firms, which initially pose little threat to more established technologies. However, the development of new compatible applications expands the market for successful new technologies over time, thereby increasing the profits earned by their producers. Rosenberg (1976) discusses a number of historical examples, where such complementarities were important. A classical example is the steam engine. This had been invented in the early part of the XVIIIth Century, but its diffusion remained very sporadic before a number of complementary innovations (e.g., Watt's separate condenser) made it competitive with the waterwheels, which remained widespread until late in the XIXth Century.

Complementarity in innovation raises interesting issues concerning the enforcement and design of intellectual property rights. For instance, what division of the surplus between basic and secondary innovation maximizes social welfare? This issue is addressed by Scotchmer and Green (1995) who construct a model where innovations are sequentially introduced, and the profits of major innovators can be undermined by subsequent derivative innovations. In this case, the threat of derivative innovations can reduce the

incentive for firms to invest in major improvements in the first place. However, too strong a defense of the property right of basic innovators may reduce the incentive to invest in socially valuable derivative innovations. Scotchmer and Green (1995) show that the optimal policy in fact consists of a combination of finite breath and length of patents. Scotchmer (1996) instead argues that it may be optimal to deny patentability to derivative innovations, instead allowing derivative innovations to be developed under licensing agreements with the owner of the basic technology. More recently, Bessen and Maskin (2002) show that when there is sufficient complementarity between innovations (as in the case of the software industry), weak patent laws may be conducive to more innovation than strong patent laws. The reason is that while the incumbent's current profit is increased by strong patent laws, its prospect of developing future profitable innovation is reduced when patent laws inhibit complementary innovations.

While this literature focuses on the partial equilibrium analysis of single industries, complementarity in innovation also has implications on broader development questions. Multiple equilibria originating from coordination failures [of the type emphasized, in different contexts, by Murphy, Shleifer and Vishny (1989), and Cooper and John (1988)] can arise when there is complementarity in innovation. Countries can get locked-in into an equilibrium with no technology adoption, and temporary big-push policies targeting incentives to adopt new technologies may turn out to be useful.[36] One such example is Ciccone and Matsuyama (1996). In their model, multiple equilibria and poverty traps may arise from the two-way causality between the market size of each intermediate good and their variety: when the availability of intermediates is limited, final good producers are forced to use a labor intensive technology which, in turn, reduces the incentive to introduce new intermediates.

Young (1993) constructs a model where innovation expands the variety of both intermediate and final goods. New intermediate inputs are not used by mature final industries, and their market is initially thin. The expansion of the market for technologies over time creates complementarity in innovation. The details of this model are discussed in the remainder of this section. To this aim, we augment the benchmark model of Section 2 with the endogenous expansion in the variety of final goods.[37] Over time, innovative investments make new intermediate inputs available to final producers, as in Romer's model. However, as a by-product (spillover), they also generate an equivalent expansion of the set of final goods that can be produced. There are no property rights defined on the production of new final goods, and these are produced by competitive firms extraneous to the innovation process.

[36] Interestingly, in models with complementarity in innovation, market economies may be stuck in no-growth traps that are inefficient in the sense that the optimal intertemporal allocation would require positive investment and growth. See, for example, Ciccone and Matsuyama (1999).

[37] Models featuring an expanding variety of final products include Judd (1985), Grossman and Helpman (1989, Chapter 3) and, more recently, Xie (1998) and Funke and Strulik (2000). Here, we follow Young (1993) which, in turn, is close to Judd's paper.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 36
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

A_t will now denote the measure of both final goods and intermediate goods available in the economy at t. Final products are imperfect substitutes in consumption, and the instantaneous utility function is:[38]

$$V_t = \int_0^{A_t} \ln(C_{s,t})\,\mathrm{d}s,$$

with total utility being

$$U = \int_0^{\infty} \mathrm{e}^{-\rho t} V_t\,\mathrm{d}t.$$

This specification implies that consumers' needs grow as new goods become available. Suppose, for instance, that a measure ε of new goods is introduced between time t and $t + j$. At time t, consumers are satisfied with not consuming the varieties yet to be invented. However, at time $t + j$, the same consumers' utility would fall to minus infinity if they did not consume the new goods.

The productive technology for the s'th final good is given by

$$Q_s = \left(\int_0^{\min[s\Theta, A]} x_{j,s}^{\alpha}\,\mathrm{d}j\right)^{1/\alpha}, \tag{59}$$

where $\Theta \geqslant 1$ is a parameter. Note that labor is not used in the final goods production. First, to build the intuition in the simplest case, we maintain that all final goods are produced with the same technology employing all available varieties of intermediate inputs. More formally, we characterize the equilibrium in the limit case where $\Theta \to \infty$, so that $\min[s\Theta, A] = A$. This assumption will be relaxed later.

We use the "labor-for-intermediates" model introduced in Section 2.2, where labor is used for research and intermediate production and the productivity of labor in intermediate production equals A_t. We choose the nominal wage as the numeraire.[39] Hence, the profit of an intermediate producer can be expressed as:

$$\pi = \frac{1-\alpha}{\alpha}\frac{x}{A} = \frac{1-\alpha}{\alpha}\frac{L - L_x}{A}. \tag{60}$$

Note that profits fall over time at the rate at which knowledge grows. In a BG equilibrium, the interest rate is constant and A grows at the constant rate γ. Free entry implies:

$$\int_t^{\infty} \mathrm{e}^{-r\tau}\pi_\tau\,\mathrm{d}\tau = \frac{1-\alpha}{\alpha}\frac{L - \gamma/\delta}{A_t(r+\gamma)} \leqslant \frac{1}{\delta A_t},$$

[38] This is the benchmark specification in Young (1993), where it is then extended to general CES preferences across goods. The logarithmic specification is analytically convenient because of the property that consumers spend an equal income share on all existing goods.

[39] Note that we cannot simply set the price of the final good as the numeraire, as there is an increasing variety of final goods.

where we have used the fact that $\gamma = \delta L_x$. Simplifying terms yields

$$\frac{1-\alpha}{\alpha}\frac{\delta L - \gamma}{r+\gamma} \leqslant 1 \tag{61}$$

The intertemporal optimality condition for consumption also differs from the benchmark model. In particular, if E denotes the total expenditure in final goods, the Euler condition is:[40]

$$\frac{\dot{E}}{E} = r - \rho + \frac{\dot{N}_t}{N_t}.$$

In a BG equilibrium, the total expenditure on consumption goods is constant. Hence,

$$r = \rho - \gamma. \tag{62}$$

This expression can be substituted into (61) to give a unique solution for γ. As long as growth is positive, we have

$$\gamma = \delta L - \frac{\alpha\rho}{1-\alpha},$$

which is almost identical to (15), except for the constant term $\alpha/(1-\alpha)$ being replaced by $1/\alpha$. In the limit case considered so far ($\Theta \to \infty$), the model is isomorphic to Romer (1990).

Next, we move to the general case where $\Theta \geqslant 1$ is finite. This implies that final producers cannot use the entire range of intermediate goods. In particular, an intermediate good indexed by s cannot be used by "mature" final industries having an index j, such that $j < s/\Theta$. This assumption captures the idea that a technology mismatch develops over time between mature final good industries and new technologies.[41]

An important implication of this assumption is that, when introduced, a new technology (intermediate input) is only required by a limited number of final industries. Thus, the monopolist producing a new variety has a small cash-flow. This is especially true when the parameter Θ is small: as $\Theta \to 1$, there is no demand for a new intermediate good at the time of its first appearance. However, the market for technologies expands over time, as new final goods using "modern" technologies appear. This dynamic market size effect generates complementarity in the innovation process. An innovator is eager to see rapid technical progress, as this expands the number of users of the new technology.

Countering this effect, there is a process of "expenditure diversion" that reduces, *ceteris paribus*, the demand for each intermediate good. Over time, technical progress expands the number of intermediate inputs over which final producers spread their

[40] See Young (1993, p. 783) for the derivation of this Euler equation.

[41] In principle, it would seem natural to assume that new final goods do not use very old intermediate goods. Young (1993, p. 780) argues that allowing for this possibility would not change the main results, but would make the analysis more involved.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 38
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

demand. As noted above, the total expenditure on final goods is constant in a BG equilibrium. Since final good firms make zero profits at all times, and intermediates are the only inputs, the total expenditure on the intermediate goods must also be constant. Therefore, an increase in A dilutes the expenditure over a larger mass of intermediate goods, and reduces the profit of each existing intermediate firm. This effect generates substitution rather than complementarity in innovation.

The dynamic market size effect may dominate for young intermediate firms. But as a technology becomes more mature, the expenditure dilution effect takes over. Thus, firms can go through a life-cycle: their profit flow increases over time at an earlier stage and decreases at a later stage.

We denote by $\pi(A_\tau, A_t)$ the profit realized at time τ by an intermediate producer who entered the market in period $t < \tau$. Solving the profit maximization problem for the intermediate monopolist, subject to the demand from final industries, leads to the following expression:

$$\pi(A_\tau, A_t) = \frac{1-\alpha}{\alpha A_\tau}(L - L_x)\left(1 + \frac{\gamma(\tau - t) - 1}{\Theta}\right). \tag{63}$$

It is easily verified that as $\Theta \to \infty$, the solution becomes identical to (60), where nominal profits fall at the same rate as A_t.

Free-entry implies:

$$\int_t^\infty \mathrm{e}^{-r(\tau-t)}\pi(A_\tau, A_t)\,\mathrm{d}\tau \leqslant \frac{1}{\delta A_t}. \tag{64}$$

Solving the integral on the left-hand side, using the Euler condition, $r + \gamma = \rho$, and simplifying terms yields the following equilibrium condition:

$$f_{\mathrm{FE}}(\gamma) = \frac{1-\alpha}{\alpha\rho^2\Theta}(\delta L - \gamma)\big(\gamma + \rho(\Theta - 1)\big) \leqslant 1, \tag{65}$$

where all terms but γ are parameters. For sufficiently large values of Θ, i.e., when the market for new technology is large, $f'_{\mathrm{FE}}(\gamma) < 0$ and the equilibrium is unique. However, if $\Theta < 1 + \delta L/\rho$, $f_{\mathrm{FE}}(\gamma)$ is non-monotonic, and multiple equilibria are possible.

Figure 2 describes the three possible cases. As long as $\rho > \gamma$, which is a necessary and sufficient condition for the interest rate to be positive, $f_{\mathrm{FE}}(\gamma)$ is increasing in Θ. For a range of small Θ's, there is no equilibrium with positive innovation (lower curve). The only equilibrium is a point such as X', featuring zero growth.

For an intermediate range of Θ, we have $f_{\mathrm{FE}}(\gamma) = 1$ in correspondence of two values of γ (intermediate curve). This implies that (for generic economies), there exist three equilibria, where equilibria such as point X feature zero innovation and growth. Firms contemplating entry expect no expansion of the market size for new technologies. Furthermore, such market size is too small to warrant profitable deviations, and the expectation of no innovation is fulfilled in equilibrium. Equilibria such as point Y are

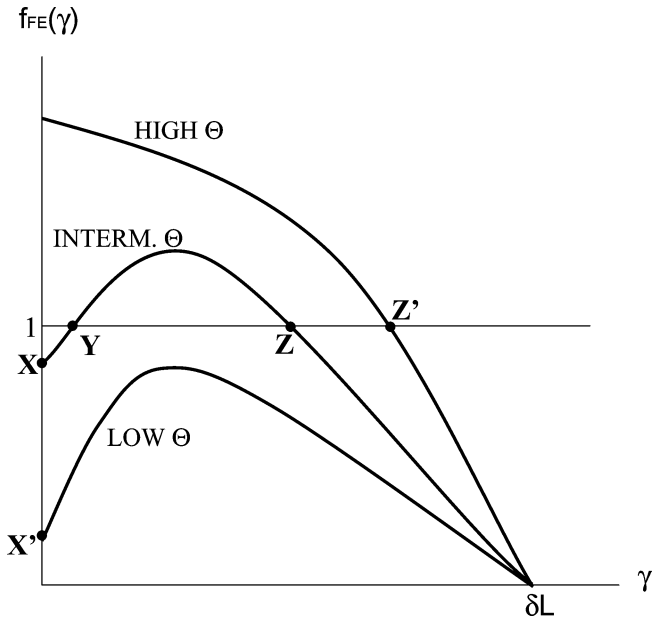


Figure 2.

characterized by local complementarity in innovation: the expectation of higher future innovation and growth increases the value of new firms, stimulating current entry and innovation. In steady-state (BG), this implies a positive slope of the locus $f_{\mathrm{FE}}(\gamma)$.[42] Eventually, for sufficiently high growth rates, the diversion effect dominates. Thus, in an equilibrium like Z, the value of innovating firms depends negatively on the speed of innovation.[43]

Finally, for a range of large Θ's, substitution dominates throughout (upper curve). The initial market for new technologies is sufficiently large to make the expenditure diversion effect dominate the market size effect, even at low growth rates. The equilibrium is unique, and the solution is isomorphic to that of the benchmark model of expanding variety.

[42] As mentioned above, firms go through a life-cycle here. When a new technology is introduced, the profit flow of an innovating firm is small. As time goes by, the expenditure diversion effect becomes relatively more important. The value of a firm upon entry is the PDV of its profit stream. Local complementarity occurs if, for a particular γ, profits increase at a sufficiently steep rate in the earlier part of the firm's life-cycle.

[43] If the expectational stability of the equilibria in the sense of Evans and Honkapohja (2001) is tested, equilibria such as point Y are not found to be E-stable, while equilibria such as X and Z are stable. See the discussion in Section 7.2.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 40
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

6. Financial development

A natural way in which the expansion of the variety of industries can generate complementarities in the growth process is through its effects on financial markets. Acemoglu and Zilibotti (1997) construct a model where the introduction of new securities, associated with the development of new intermediate industries, improves the diversification opportunities available to investors. Investors react by supplying more funds, which fosters further industrial and financial development, generating a feedback.[44] The model offers a theory of development. At early stages of development, a limited number of intermediate industries are active (due to technological nonconvexities), which limits the degree of risk-spreading that the economy can achieve. To avoid highly risky investments, agents choose inferior but safer technologies. The inability to diversify idiosyncratic risks introduces a large amount of uncertainty in the growth process. In equilibrium, development proceeds in stages. First, there is a period of "primitive accumulation" with a highly variable output, followed by take-off and financial deepening and finally, steady growth. Multiple equilibria and poverty traps are possible in a generalized version of the model.

The theory can explain why the growth process is both slow and highly volatile at early stages of development, and stabilizes as an economy grows richer. Evidence of this pattern can be found in the accounts of pre-industrial growth given by a number of historians, such as Braudel (1979), North and Thomas (1973) and DeVries (1990). For instance, in cities such as Florence, Genoa and Amsterdam, prolonged periods of prosperity and growth have come to an end after episodes of financial crises. Interestingly, these large set-backs were not followed (as a neoclassical growth model would instead predict) by a fast recovery but, rather, by long periods of stagnation. Similar phenomena are observed in the contemporary world. Acemoglu and Zilibotti (1997) document robust evidence of increases in GDP per capita being associated with large decreases in the volatility of the growth process. It has also been documented that higher volatility in GDP is associated with lower growth [Ramey and Ramey (1995)].

We here describe a simplified version of the model. Time is discrete. The economy is populated by overlapping generations of two-period lived households. The population is constant, and each cohort has a unit mass ($L = 1$). There is uncertainty in the economy, which we represent by a continuum of equally likely states $s \in [0, 1]$. Agents are assumed to consume only in the second period of their lives.[45] Their preferences are parameterized by the following (expected) utility function, inducing unit relative risk

[44] This paper is part of a recent literature on the two-way relationship between financial development and growth. This includes Bencivenga and Smith (1991), Greenwood and Jovanovic (1990) and Zilibotti (1994). In none of these other papers does financial development take the form of an expansion in the "variety" of assets.

[45] This is for simplicity. Acemoglu and Zilibotti (1997) assume that agents consume in both periods. It is also possible to study the case of a general CRRA utility function.

aversion:

$$E_t U(c_{t+1}) = \int_0^1 \log\bigl(c_{t+1}^s\bigr)\mathrm{d}s. \tag{66}$$

The production side of the economy consists of a unique final good sector, and a continuum of intermediate industries. The final good sector uses intermediate inputs and labor to produce final output. Output in state s is given by the following production function:

$$Y_{s,t} = (x_{s,t-1} + x_{\Phi,s,t-1})^{\alpha} L^{1-\alpha}. \tag{67}$$

The term in brackets is "capital", and it is either produced by a continuum of intermediate industries, each producing some state-contingent amount of output (x_s), or a separate sector using a "safe technology" (x_Φ). The measure of the industries with a state-contingent production, A_t, is determined in equilibrium, and A_t can expand over time, like in Romer's model, but it can also fall. Moreover, $A_t \in [0, 1]$, i.e., the set of inputs is bounded.

In their youth, agents work in the final sector and earn a competitive wage, $w_{s,t} = (1-\alpha)Y_{s,t}$. At the end of this period, they take portfolio decisions: they can place their savings in a set of risky securities ($\{F_i\}_{i\in[0,A_t]}$), consisting of state-contingent claims to the output of the intermediate industries or, in a safe asset (ϕ), consisting of claims to the output of the safe technology. After the investment decisions, the uncertainty unravels, the security yields its return and the amount of capital brought forward to the next period is determined. The capital is then sold to final sector firms and fully depreciates after use. Old agents consume their capital income and die.

Intermediate industries use final output for production. An intermediate industry $i \in [0, A_t]$ is assumed to produce a positive output only if state $s = i$ occurs. In all other states of nature, the firm is not productive. Moreover, the ith industry is only productive if it uses a minimum amount of final output, M_i, where

$$M_i = \max\left\{0, \frac{D}{(1-x)}(i-x)\right\},$$

with $x \in (0, 1)$. This implies that some intermediate industries require a certain minimum size, M_i, before being productive. In particular, industries $i \leqslant x$ have no minimum size requirement, and for the rest of the industries, the minimum size requirement increases linearly with the index i.

To summarize, the intermediate technology is described by the following production function:

$$x_{i,s} = \begin{cases} RF_i & \text{if } i = s \text{ and } F_i \geqslant M_i, \\ 0 & \text{otherwise.} \end{cases}$$

Since there are no start-up costs, all markets are competitive. Thus, firms retain no profits, and the product is entirely distributed to the holders of the securities. The jth security entitles its owner to a claim to R units of capital in state j (as long as the minimum size constraint is satisfied, which is always the case in equilibrium), and otherwise

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 42
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

to nothing. Savings invested in the "safe technology" give the return

$$x_{\Phi,s} = r\phi, \quad \forall s \in [0, 1],$$

where $r < R$. Thus, one unit of the safe asset is a claim to r units of capital in all states of nature.

Since the risky securities yield symmetric returns, and there is safety in numbers, it is optimal for risk-averse agents to hold a portfolio containing all available securities in equal amounts. More formally, the optimal portfolio decision features $F_i = F$, for all $i \in [0, A_t]$. We refer to this portfolio consisting of an equal amount of all traded risky securities as a *balanced portfolio*.

If $A_t = 1$, a balanced portfolio of risky securities bears no risk, and first-order dominates the safe investment. However, due to the presence of technological non-convexities (minimum size requirements), not all industries are in general activated. When $A_t < 1$, the inferior technology is safer, and there is a trade-off between risk and productivity. In this case, the optimal investment decision of the representative saver can be written as:

$$\max_{\phi_t, F_t} A_t \log\left[\rho_{G,t+1}(RF_t + r\phi_t)\right] + (1 - A_t)\log\left[\rho_{B,t+1}(r\phi_t)\right], \tag{68}$$

subject to

$$\phi_t + A_t F_t \leqslant w_t. \tag{69}$$

$\rho_{\tilde{s},t+1}$ denotes the rate of return of capital, which is taken as parametric by agents, and does not affect the solution of the program.[46] Agents also take A_t, i.e., the set of securities offered, as parametric.

Simple maximization yields:

$$\phi_t^* = \frac{(1 - A_t)R}{R - rA_t} w_t, \tag{70}$$

$$F_{i,t}^* = \begin{cases} F(A_t) \equiv \frac{R-r}{R-rA_t} w_t, & \forall i \leqslant A_t, \\ 0 & \forall i > A_t. \end{cases} \tag{71}$$

Figure 3 expresses the demand for each risky asset, $F(A_t)$ (FF schedule), as a function of the measure of intermediate industries which are active. The FF schedule is

[46] In equilibrium:

$$\rho_{G,t+1} = \alpha(RF_t + r\phi_t)^{\alpha-1}$$

and

$$\rho_{B,t+1} = \alpha(r\phi_t)^{\alpha-1}.$$

$\rho_{G,t+1}$ applies in the "good state", i.e., when the realized state is $i \leqslant A_t$, while $\rho_{B,t+1}$ is the marginal product of capital in the "bad" state, when the realized state is $i > A_t$ and no risky investment pays off.

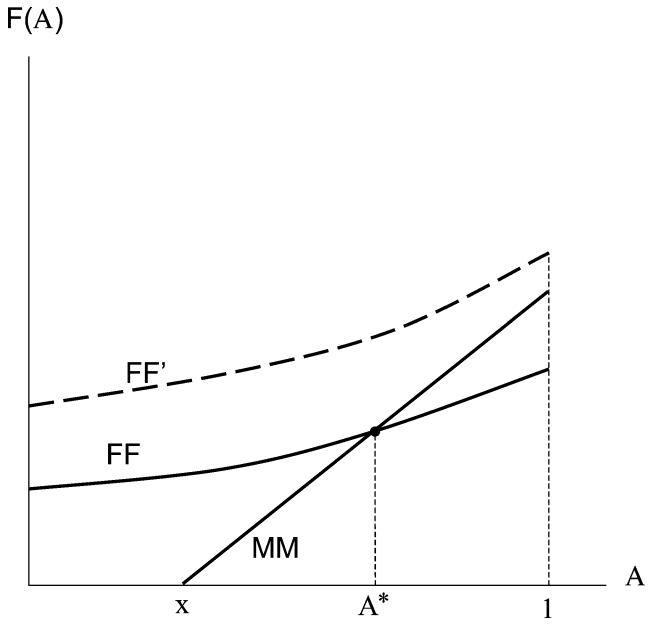


Figure 3.

upward sloping, implying that there is complementarity in the demand for risky assets: the demand for *each asset* grows with the variety of intermediate industries.

Complementarity arises because the more active are intermediate industries, the better is risk-diversification. Thus, as A_t increases, savers shift their investments away of the safe asset into high-productivity risky projects (the "stock market"). Such complementarity hinges on risk aversion being sufficiently high.[47] In general, similar to Young (1993), an increase in A creates two effects. On the one hand, investments in the stock market become safer because of better diversification opportunities, which induces complementarity. On the other hand, investments are spread over a larger number of assets, inducing substitution. With sufficiently high risk aversion, including the unit CRRA specification upon which we focus, the first effect dominates.

The equilibrium measure of active industries, A_t^*, is determined (as long as $A^* < 1$) by the following condition:

$$F\left(A_t^*\right) = M_{A_t^*}.$$

[47] Suppose agents were risk averse, but only moderately so. Suppose, in particular, that they were so little risk-averse that they would decide not to hold any safe asset in their portfolio. Then, an expansion in the set of risky securities would induce agents to spread their savings (whose total amount is predetermined) over a larger number of assets. In this case, assets would be substitutes rather than complements.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 44
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

In Figure 3, the equilibrium is given by the intersection of schedules FF and MM, where the latter represents the distribution of minimum size requirements across industries. Intuitively, A_t^* is the largest number of industries for which the technological non-convexity can be overcome, subject to the demand of securities being given by (71).[48]

Growth increases wage income and the stock of savings over time. In equilibrium, this induces an expansion of the intermediate industries, A_t^*. This can once more be seen in Figure 3: growth creates an upward shift of the FF schedule, causing the equilibrium to move to the left. Therefore, growth triggers financial development. In particular, when the stock of savings becomes sufficiently large, the financial market is sufficiently thick to allow all industries to be active. In the case described by the dashed curve, FF′, the economy is sufficiently rich to afford $A_t^* = 1$. The inferior safe technology is then abandoned. Financial development, speeds up growth by channelling investments towards the more productive technology.

The stochastic equilibrium dynamics of GDP can be explicitly derived:

$$Y_{t+1} = \begin{cases} F_B(Y_t) = \left((1-\alpha)\dfrac{r(1-A_t^*)}{R - rA_t^*} RY_t\right)^{\alpha} & \text{prob. } 1 - A_t^*, \\ F_G(Y_t) = \left((1-\alpha)RY_t\right)^{\alpha} & \text{prob. } A_t^*, \end{cases} \tag{72}$$

where $A_t^* = A(Y_{t-1}) \leqslant 1$ is the equilibrium measure of intermediate industries, such that $A' \geqslant 0$.[49] The first line corresponds to the case of a "bad realization" at time t, such that $s \in (A_t^*, 1]$. In this case, none of the active intermediate industries turned out to pay-off at time t, and capital at time $t+1$ is only given by the return of the safe technology. The second line corresponds to the case of a "good realization" at t, such that $s \in [0, A_t^*]$. In this case, the risky investment paid off at time t, and capital and output are relatively large at time $t+1$. Note that the probability of a good realization increases with the level of development, since $A'(Y_{t-1}) \geqslant 0$ (with strict inequality as long as $A^* < 1$).

Figure 4 describes the dynamics. The two schedules represent output at time $t+1$ as a function of output at time t conditional on good news ($F_G(Y_t)$) and bad news ($F_B(Y_t)$), respectively. At low levels of capital ($Y \leqslant Y_L$), the marginal product of capital is very high, which guarantees that growth is positive, even conditional on bad news. In the intermediate range where $Y \in [Y_L, Y_M]$, growth only occurs if news is good, since $F_B(Y_t) < Y_t < F_G(Y_t)$. The threshold Y_L is not a steady-state; however, it is a point around which the economy will spend some time. When the initial output is below Y_L, the economy necessarily grows towards it. When it is above Y_L, output falls back whenever bad news occurs. So, in this region, the economy is still exposed to undiversified risks, and experiences fluctuations and set-backs. Finally, for $Y \geqslant Y_M$, there are

[48] Acemoglu and Zilibotti (1997) show the laissez-faire portfolio investment to be inefficient. Efficiency would require more funds to be directed to industries with large non-convexities, i.e., agents not holding a balanced portfolio. The inefficiency is robust to the introduction of a rich set of financial institutions.

[49] Acemoglu and Zilibotti (1997) derive a closed-form solution for A_t^* that we do not report here.

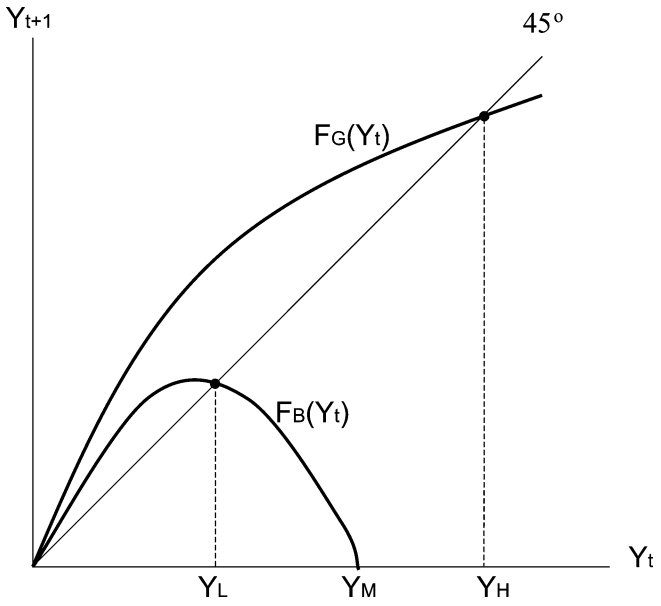


Figure 4.

enough savings in the economy to overcome all technological non-convexities. When the economy enters this region, all idiosyncratic risks are removed, and the economy deterministically converges to Y_H.[50]

Note that it may appear as if, in the initial stage, countries striving to take off do not grow at a sustained rate during long periods. The demand for insurance takes the form of investments in low-productivity technologies, and poor economies tend to have low total factor productivity and slow growth.

In the case described by Figure 4, the economies "almost surely" converge to a unique steady-state. Different specifications of the model can, however, lead to less optimistic predictions. With higher risk aversion, for instance, traps can emerge, as in the example described in Figure 5. An economy starting with a GDP in the region $[0, Y_{MM})$ would never attain the high steady-state Y_H, and would instead perpetually wander in the trapping region $[0, Y_{LL}]$. Conversely, an economy starting above Y_M would certainly converge to the high steady-state, Y_H. Finally, the long-run fate of an economy starting in the region $[Y_{MM}, Y_M]$ would be determined by luck: an initial set of positive

[50] That the economy converges "almost surely" to a steady-state where all risk is diversified away only occurs under parameter restrictions ensuring that $Y^{SS} > Y^1$. Although the model presented here is neoclassical and features zero growth in the long run, it is possible to augment it with spillover of the learning-by-doing type, as in Romer (1986), and make it generate self-sustained growth.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 46
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

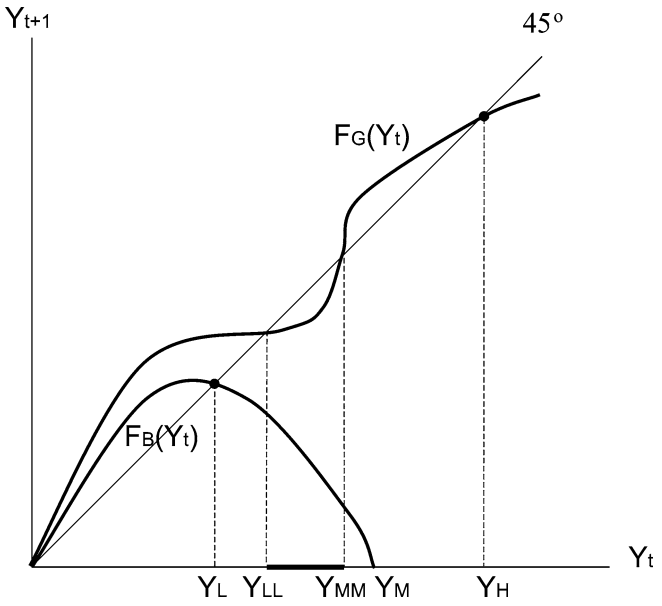


Figure 5.

draws would bring this economy into the basin of attraction of the good equilibrium. A single set-back, however, would forever jeopardize its future development.[51]

The model can be extended in a number of directions. A two-country extension shows that international capital flows may lead to divergence, rather than convergence between economies. This result is due to the interplay between two forces: first, decreasing returns to capital would tend to direct foreign investments towards poorer countries, as in standard neoclassical models. Second, the desire to achieve better diversification pushes investments towards thicker markets. The latter force tends to prevail at some earlier stages of the development process. So, poor countries suffer an outflow of capital, which spills over to lower income and wages for the next generation, thereby slowing down the growth process. The analysis of capital flows, financial integration and financial crises in the context of similar models is further developed in recent papers by Martin and Rey (2000, 2001 and 2002). A different extension of the model is pursued by Cetorelli (2002) who shows that the theory can account for phenomena such as "club convergence", economic miracles, growth disasters and reversals of fortune.

[51] Consider, for instance, the limit case where agents are infinitely risk-averse. In this case, agents refuse to invest in the stock-market as long as this entails some uncertainty, i.e., as long as there are not enough savings in the economy to open all industries. Thus, an economy starting above Y_M converges to Y_H, while an economy starting below Y_M converges to Y_L, and is stuck in a poverty trap.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 47
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

Recent empirical studies analyze implications of the theory about the patterns of risk-sharing and diversification. Kalemli-Ozcan et al. (2001, 2003) document that regions with access to better insurance through capital markets can afford a higher degree of specialization. Using cross-country data at different levels of disaggregation, Imbs and Wacziarg (2003) find robust evidence of sectoral diversification increasing in GDP. However, their findings also suggest that, at a relatively late stage of the development process, the pattern reverts and countries once more start to specialize. This tendency for advanced countries to become more specialized as they grow can be explained by factors emphasized by the "new economic geography" literature [Krugman (1991)], from which the theory described in this section abstracts, such as agglomeration externalities and falling transportation costs.

7. Endogenous fluctuations

In the models reviewed so far, the economies converge in the long run to balanced-growth equilibria characterized by linear dynamics. Growth models with expanding variety and technological complementarities can, however, generate richer long-run dynamics, including limit cycles. In this section, we review two such models.

In the former, based on Matsuyama (1999), cycles in innovation and growth arise from the deterministic dynamics of two-sector models with an endogenous market structure. The theory can explain some empirical observations about low-frequency cycles, and their interplay with the growth process. In particular, it predicts that waves of rapid growth mainly driven by "factor accumulation" are followed by spells of innovation-driven growth. Interestingly, these latter periods are characterized by lower investments and slower growth. This is consistent with the findings of Young (1995) that the growth performance of East-Asian countries was mainly due to physical and human capital accumulation, while there was little total factor productivity (TFP) growth. According to Matsuyama's theory, the observation of low TFP growth should not lead to the pessimistic conclusions that growth is destined to die-off. Rather, rapid factor accumulation could set the stage for a new phase of growth characterized by more innovative activity. The predictions of this theory bear similarities to those of models with General Purpose Technologies (GPT), e.g., Helpman (1998) and Aghion and Howitt (1998, Chapter 8). For instance, they predict that a period of rapid transformation and intense innovation (e.g., the 1970's) can be associated with productivity slowdowns. However, GPT-based theories rely on the exogenous arrival of new "fundamental" innovations generating downstream complementarities. In contrast, cycles in Matsuyama (1999) are entirely endogenous.[52]

[52] Cyclical equilibria can also emerge in Schumpeterian models, due to the dynamic relationship between innovative investments and creative distruction. An example is the seminal contribution of Aghion and Howitt (1992). More recently, Francois and Lloyd-Ellis (2003) construct a Schumpeterian model where entrepreneurs

In the latter model, based on Evans et al. (1998), cycles in innovation and growth are instead driven by expectational indeterminacy. The mechanism in this paper is different, as cycles hinge on multiple equilibria and sunspots. Some main predictions are also different: contrary to Matsuyama (1999), the equilibrium features a positive comovement of investments and innovation. The main contribution of the paper is to show that cycles can be learned by unsophisticated agents holding adaptive expectations. Thus, the predictive power of the theory does not rest on the assumption that agents' expectations are rational and that agents can compute complicated dynamic equilibria.

7.1. Deterministic cycles

Matsuyama (1999) presents a model of expanding variety where an economy can perpetually oscillate in equilibrium between periods of innovation and periods of no innovation. Cycles arise from the deterministic periodic oscillations of two state variables (physical capital and knowledge). Unlike the model that will be discussed in the next section, the equilibrium is determinate and there are no multiple steady-states.

More specifically, the source of the oscillatory dynamics is the market structure of the intermediate goods market. Monopoly power is assumed to be eroded after one period. The loss of monopoly power is due to the activity of a competitive fringe which can copy the technology with a one-period lag. In every period, new industries are monopolized, while mature industries are competitive. The profits of innovators depend on the market structure of the intermediate sector. The larger is the share of competitive industries in the intermediate sector, the lower is the profit of innovative firms, since competitive industries sell larger quantities and charge lower prices. In periods of high innovation, a large share of industries are monopolized, which increases the profitability of innovation, thereby generating a feedback. In these times, investment in physical capital is low due to the crowding out from the research activity. Conversely, times of low innovation are times of high competition, since old monopolies lose power and there are few new firms. Thus, the rents accruing to innovative firms are small. In these periods, savings are invested in physical capital, and while innovation is low, the high accumulation of physical capital creates the conditions for future innovation to be profitable.

Time is discrete. The production of final goods is as in (3), where we set $L_y = L = 1$. Intermediate goods are produced using physical capital, with one unit of capital producing one unit of intermediate product, x. Innovation also requires capital, with a requirement of μ units of capital per innovation. Monopoly power is assumed to last one period only. Therefore, in period t, all intermediate inputs with an index $z \in [0, A_{t-1}]$

can decide to time the implementation of innovations [similarly to Shleifer (1986)]. In this model, agents time the implementation so as to profit from buoyant demand and maximize the duration of their leadership. This mechanism leads to a clustering of innovations and endogenous cycles. While this model can explain some features of fluctuations at business cycle frequencies, Matsuyama's model is better suited for the analysis of long waves.

are competitively priced, whereas all those with an index $z \in [A_{t-1}, A_t]$ are monopolistically priced. The prices of competitive and non-competitive varieties are $p_t^c = r_t$ and $p_t^m = r_t/\alpha$, respectively, where the superscript $h \in \{c, m\}$ denotes the market structure. The relative demand for two varieties x_t^c and x_t^m must be

$$\frac{x_t^c}{x_t^m} = \left(\frac{p_t^c}{p_t^m}\right)^{-1/(1-\alpha)} = \alpha^{-1/(1-\alpha)}. \tag{73}$$

The one-period monopoly profit is $\pi_t = p_t^m x_t^m - r_t x_t^m = x_t^m r_t(1-\alpha)/\alpha$. Since patents expire after one period, π_t is also the value of a monopolistic firm at the beginning of period t. Therefore, free-entry implies:

$$\frac{1-\alpha}{\alpha} x_t^m \leqslant \mu, \tag{74}$$

with equality holding when innovation is positive.

Capital is assumed to fully depreciate after each period. The stock of capital can be allocated to research or intermediate production, subject to the following resource constraint:

$$K_{t-1} = A_{t-1} x_t^c + (A_t - A_{t-1})\big(x_t^m + \mu\big),$$

implying

$$A_t - A_{t-1} = A_{t-1} \max\left\{0, \frac{(1-\alpha)K_{t-1}}{\mu A_{t-1}} - \alpha^{-\alpha/(1-\alpha)}\right\}, \tag{75}$$

where we have used (73) and (74) to eliminate x_t^c and x_t^m. As shown by (75), there exists a threshold to the capital-knowledge ratio that triggers positive innovation. In particular, innovation occurs if $K_{t-1}/A_{t-1} \geqslant \alpha^{-\alpha/(1-\alpha)}(1-\alpha)^{-1}\mu \equiv k_L$. If $K_{t-1}/A_{t-1} < k_L$, then, all capital is allocated to intermediate production, all intermediate industries are competitive and final production is given by the standard neoclassical Cobb–Douglas technology:

$$Y_t = A_{t-1}^{1-\alpha} K_{t-1}^{\alpha}. \tag{76}$$

In this case, an economy is said to be in a "Solow regime", with decreasing returns to capital. Since there is no investment in innovation, A is constant and the dynamics has a neoclassical character. In contrast, if $K_{t-1}/A_{t-1} \geqslant k_L$, then a positive share of the capital stock is allocated to innovation and final production equals:

$$Y_t = A_{t-1}\left[\alpha^{-1/(1-\alpha)}\frac{\alpha}{1-\alpha}\mu\right]^{\alpha} + (A_t - A_{t-1})\left[\frac{\alpha}{1-\alpha}\mu\right]^{\alpha}.$$

Using (75) and simplifying terms, this equation can be written as

$$Y_t = DK_{t-1}, \tag{77}$$

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 50
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

where $D \equiv (k_L)^{-(1-\alpha)}$. In this case, the returns to capital are constant, like in endogenous growth models, and the economy is said to be in a "Romer regime".[53]

For tractability, we assume a constant savings rate, implying that $K_t = sY_t$.[54] Define $k_t = k_L^{-1} \cdot K_t/A_t$ as the (adjusted) capital-to-knowledge ratio. Then, standard algebra using (75), (76) and (77) establishes the following equilibrium law of motion:

$$k_t = f(k_{t-1}) = \begin{cases} sDk_{t-1}^{\alpha} & \text{if } k_{t-1} < 1, \\ \dfrac{sDk_{t-1}}{1 + \alpha^{-\alpha/(1-\alpha)}(k_{t-1} - 1)} & \text{if } k_{t-1} \geqslant 1. \end{cases} \tag{78}$$

The mapping $k_t = f(k_{t-1})$ has two fixed points. The first is $k = 0$, the second can either be $k = (sD)^{1/(1-\alpha)} \equiv \hat{k}_1$, if $sD \leqslant 1$, or $k = 1 + \alpha^{\alpha/(1-\alpha)}(sD - 1) \equiv \hat{k}_2$, if $sD \geqslant 1$. In the former case, the fixed point lies in the range of the " Solow regime", while in the latter, it lies in the range of the "Romer regime".[55]

Three cases are possible:

1. If $sD \leqslant 1$, the economy converges monotonically to $k = \hat{k}_1$. In this case, the economy never leaves the Solow regime, and there are no innovative investments. The neoclassical dynamics converge to a stagnating level of GDP per capita.
2. If $sD > \max\{1, \alpha^{-\alpha/(1-\alpha)} - 1\}$, then capital first monotonically accumulates in the Solow regime, with no innovation. The economy overcomes the development threshold, $k = 1$ in finite time, and the process of innovation starts thereafter. Eventually, the economy converges to the BG equilibrium $\hat{k}_2$ in an oscillatory fashion. In the BG, capital and knowledge are accumulated at the same positive rate, and income per capita grows over time.
3. If $sD \in (1, \alpha^{-\alpha/(1-\alpha)} - 1]$, the economy does not converge asymptotically to any BG equilibrium, and perpetually oscillates in the long run between the Solow- and the Romer-regime. This case is described by Figure 6. On the one hand, there is no steady-state in the Solow-regime, which rules out that the economy can be trapped in a stable equilibrium with no innovation. On the other hand, the steady-state $\hat{k}_2$ is locally unstable and cannot be an attractor of the dynamics in itself. Instead, there exists a period-2 cycle, such that one of the periodic points lies in the Solow regime (k_S), while the other lies in the Romer regime (k_R).[56] The period-2

[53] Zilibotti (1995) finds similar dichotomic equilibrium dynamics in a one-sector model with learning-by-doing spillovers. Economies may converge to a stationary steady-state with "Solow dynamics" or embark on a virtuous path of "Romer dynamics" with self-sustained growth. Cycles cannot arise in equilibrium, while multiple self-fulfilling prophecies exist.

[54] Matsuyama (2001) relaxes this restriction and characterizes equilibrium by a second-order difference equation. Some of the main results, like the existence of a period-2 cycle, survive this generalization.

[55] It is easily verified that $f'(0) > 1$, $f'(\hat{k}_1) = \alpha \in (0, 1)$, and $f'(\hat{k}_2) = -(\alpha^{-\alpha/(1-\alpha)} - 1)/(sD)$, where $f'(\hat{k}_2) \in (-1, 0)$ if $sD > \alpha^{-\alpha/(1-\alpha)} - 1$ and $f'(\hat{k}_2) < -1$ if $sD \in (1, \alpha^{-\alpha/(1-\alpha)} - 1)$. These properties are used to establish the results discussed below.

[56] A period-2 cycle exists if, given a mapping $x_{t+1} = f(x_t)$, $f(f(\cdot))$ has fixed points other than the fixed point of $f(\cdot)$. A sufficient condition is that (i) $f(\cdot)$ is continuous, (ii) there exists a closed, finite interval, I, such that $f(I) \subset I$ and (iii) $f(\cdot)$ has an unstable fixed point. (i) and (iii) are clearly satisfied; (iii) is established in the next footnote for the interval I_{abs}.

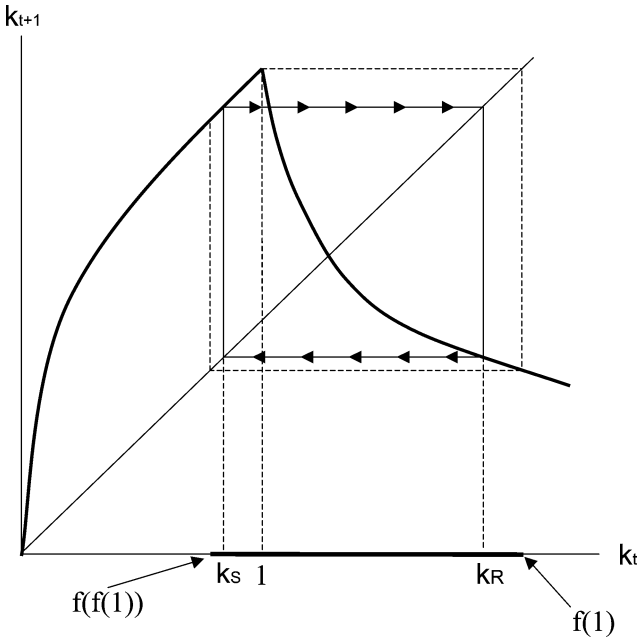


Figure 6.

cycle is not necessarily stable, and if it is unstable, the economy can converge to cycles of higher periodicity or feature chaos. A general property of the dynamics is that the economy necessarily enters the region $I_{\text{abs}} = [f(f(1)), f(1)]$ (shaded in Figure 6), and never escapes from it.[57]

In the case described by Figure 6, the model predicts that a poor economy would first grow through capital accumulation, and eventually enter the absorbing region I_{abs}. Then, there is an alternance of periods of innovation and periods of no innovation. GDP per capita grows on average, but at a non-steady rate, and there are cycles in the innovative activity. Interestingly, output and capital grow more quickly in periods of no innovation (Solow regime) than in periods of high innovation (Romer regime). Another implication is that if an economy grows quickly, but has a low TFP growth, this does not imply that growth will die-off. Rather, fast capital accumulation can create the conditions for future waves of innovation, and vice versa.

[57] To prove this result, two properties of the mapping need to be shown [see Azariadis (1993)]. First, $f(\cdot)$ must be unimodal, i.e., (i) $f(\cdot)$ must be continuous; (ii) $f(\cdot)$ must be increasing in some left-hand neighborhood of 1 and decreasing in some right-hand neighborhood of 1. Second, it must be the case that $f(f(1)) < 1$. That f is unimodal is immediate by inspection. After some algebra, it can also be proved that $f(f(1)) < 1$.

7.2. Learning and sunspots

Evans et al. (1998) propose the following generalization of the technology (3) for final production:

$$Y = L^{1-\alpha}\left[\int_0^A x_j^{\zeta}\,\mathrm{d}j\right]^{\phi}, \tag{79}$$

where $\zeta\phi = \alpha$. This specification encompasses the technology (3), in the case of $\phi = 1$, and allows intermediate inputs to be complements or substitutes. They focus on the case of complementarity ($\phi > 1$), and show that in this case, the equilibrium can feature multiple steady states, expectational indeterminacy and sunspots. They emphasize the possibility of equilibria where the economy can switch stochastically between periods of high and low growth.

Time is discrete, and intermediate firms rent physical capital from consumers to produce intermediate goods. One unit of capital is required per unit of intermediate good produced. Capital is assumed not to depreciate. The resource constraint of this economy is:

$$Y_t = C_t + K_t \cdot \chi\left(\frac{K_{t+1} - K_t}{K_t}\right),$$

where $\chi(.)$ is a function such that $\chi' > 0$, $\chi'' \geqslant 0$. If there are no costs of adjustment, then, $\chi(x) = x$. If $\chi'' > 0$, there are convex costs of adjustments.

By proceeding as in Section 2, we can characterize the equilibrium of the intermediate industry.[58] The profit of intermediate producers, in particular, turns out to be:

$$\pi = \Omega A^{\xi}(r\ p_K)^{\alpha/(\alpha-1)}, \tag{80}$$

where $\xi \equiv (\phi - 1)/(1 - \alpha)$ and $\Omega \equiv (1 - \zeta)\zeta^{(1+\alpha)/(1-\alpha)}\phi^{1/(1-\alpha)}L$ are two positive constant. We denote by p_K the relative price of capital, expressed in terms of the consumption good numeraire. If there are no adjustment costs, then, $p_K = 1$ while, in general, $p_K = \chi'(.)$. Note that profits increase with A, as long as $\phi > 1$.

Two technical assumptions ensure that the model has BG properties. First, the design of a new good requires A_t^{ξ} units of output. Second, innovative investments incurred at t only give the first profit in period $t + 1$. Free entry then implies:

$$\sum_{s=0}^{\infty} \frac{\pi_{t+s}}{(1+r)^{s+1}} \leqslant p_{K,t} A_t^{\xi}. \tag{81}$$

In a BG equilibrium, consumption and capital grow at the common rate, γ. When $\phi > 1$, this rate exceeds the growth rate of technical knowledge, $\gamma_A \equiv A_{t+1}/A_t$. In particular,

[58] Note that firms rent, and do not own, their capital stock. Adjustment costs are borne at the aggregate level, not at the level of each decision unit. Therefore, it continues to be legitimate to write the profit maximization problem for intermediate producers as a sequence of static maximization problems.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 53
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

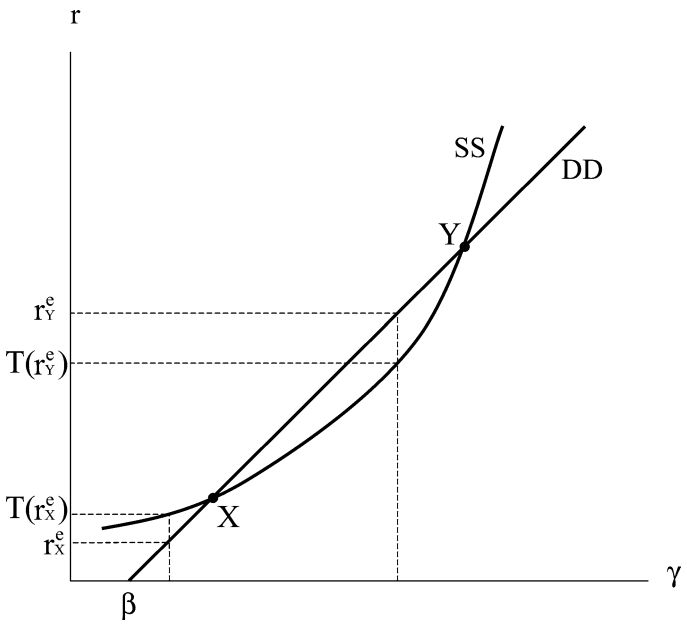


Figure 7.

it can be shown that $\gamma = \gamma_A^{1+\xi}$. Substituting (80) into (81), and solving, yields:

$$\gamma = \left(1 + r - \Omega(p_K)^{-1/(1-\alpha)} r^{-\alpha/(1-\alpha)}\right)^{(\phi-\alpha)/(\phi-1)}, \tag{82}$$

which is the analogue of Equation (13) in the benchmark model.

The model is closed by the (discrete time) Euler equation for consumption:

$$\gamma = \left[\beta(1+r)\right]^{1/\sigma}, \tag{83}$$

where β is the discount factor. Equations (82) and (83) fully characterize the equilibrium.

Figure 7 provides a geometric representation for the case of logarithmic preferences and zero adjustment cost ($p_K = 1$). The SS curve is linear, with the slope β^{-1}. The DD curve is also positively sloped. In the case represented, the two curves cross twice, thereby implying that there are two BG equilibria featuring positive innovation and growth (points X and Y).

Standard stability analysis is inappropriate for dynamic models with perfect foresight. It is possible, however, to analyze the expectational stability (E-stability) of the BG equilibria. E-stability is tested as follows. Set an arbitrary initial level for the expected interest rate r^e, and let agents choose their optimal savings plan according to (83). This implies a notional growth rate of consumption and capital, as determined by the SS curve. Next, firms take action. At the notional growth rate, there is a unique interest

rate consistent with the no-arbitrage condition implied by (82), as shown by the DD curve. The composition of these two operations define a mapping from an expected to a realized interest rate:

$$r_t = T\left(r_t^e\right). \tag{84}$$

A perfect foresight BG equilibrium is a fixed point to the mapping, $r = T(r)$. After consumers have observed the realized interest rate, they update their expectations about next period's interest rate using adaptive learning, i.e.:

$$r_{t+1}^e = r_t^e + \psi_t\left(r_t - r_t^e\right), \tag{85}$$

where $\xi_t = \psi/t$. The sequence $\{\psi_t\}$ determines how sensitive the expectations are to past errors, and it is known as the gain sequence. Substituting (84) into (85) defines a dynamic system, whose stability can be analyzed by linearization techniques. In general, expectational stability occurs whenever $T'(r) < 1$, where r is the steady-state interest rate.[59]

An inspection of Figure 7 shows the equilibrium X to be E-stable, while the equilibrium Y is not. Let r_X^e and r_Y^e denote two expected interest rates which are below the equilibria X and Y, respectively. Then, in the case of the equilibrium X, $T(r_X^e) > r_X^e$, and the adaptive adjustment moves the economy towards the equilibrium, inducing convergence. In contrast, in the case of the equilibrium Y, $T(r_Y^e) < r_Y^e$, and the adaptive adjustment moves the economy away from the equilibrium, thereby inducing divergence.

In the case analyzed so far, only one BG is E-stable, and E-stability can be used as a selection criteria. It is possible, however, that multiple E-stable BG equilibria exist in the general model with convex adjustment costs.

Figure 8 describes a case with four steady-states, two of them being E-stable. Equilibria such as X and Z are E-stable (note that $T(r_X^e) > r_X^e$ and $T(r_Z^e) > r_Z^e$). Moreover, in the neighborhood of these equilibria, there exist stationary sunspot equilibria. In one such equilibrium, the economy switches stochastically between two points in the neighborhood of X and Z, respectively, with switching probabilities given by a time-invariant transition probability matrix. The fact that both X and Z are E-stable is sufficient for any stationary sunspot equilibrium in their neighborhood to be E-stable in itself.[60]

We conclude that a modified version of the model of growth with expanding variety can generate endogenous fluctuations. The key assumptions are complementarity between capital goods and convex adjustment costs to capital. The former assumption guarantees the existence of multiple BG equilibria, around which sunspot equilibria can be constructed. The latter assumption guarantees that the sunspot equilibrium is expectationally stable, i.e., it can be learned through adaptive expectations.

[59] See Evans and Honkapohja (2001) for a state-of-art analysis of expectational indeterminacy.

[60] For general discussion of sunspot equilibria, see Azariadis and Guesnerie (1986), Grandmont (1986) and Azariadis (1993).

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 55
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

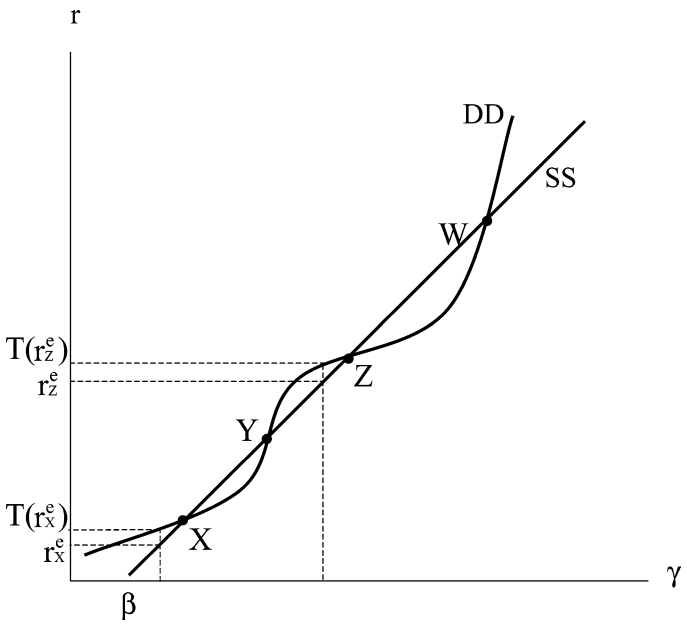


Figure 8.

The model assumes increasing returns to physical and knowledge capital. The reduced form representation of the final good technology is:

$$Y = A^{\phi-\alpha} K_F^{\alpha},$$

where $K_F = Ax$ denotes aggregate capital used in intermediate production. Empirical estimates suggest that $\alpha \leqslant 0.4$, which implies that the lower bound to the output elasticity of knowledge to generate multiplicity is $\phi - \alpha = 0.6$. Evans et al. (1998) provide a numerical example of an E-stable sunspot equilibrium, assuming $\phi = 4$. Recent estimations from Porter and Stern (2000) using patent numbers report $\phi - \alpha$ to be around 0.1, however. Therefore, the model seems to require somehow extreme parameters to generate endogenous fluctuations.

Augmenting the model with other accumulated assets, such as human capital, may help obtain the results under realistic parameter configurations. This is complicated by the presence of scale effects in the expanding variety model. However, in a recent paper, Dalgaard and Kreiner (2001) formulate a version of the model with human capital accumulation and without scale effects. In their model, both human capital (embodied knowledge) and technical change (disembodied knowledge) are used to produce final goods. The scale effect is avoided by congestion effects in the accumulation of human capital. An interesting feature of this model is that, unlike other recent models without scale effects, positive long-run growth in income per capita does not hinge on positive population growth.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 56
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

8. Conclusions

In this chapter, we have surveyed recent contributions to growth theory inspired by Romer's (1990) expanding variety model. Key features of the theory are increasing returns through the introduction of new products that do not displace existing ones and the existence of monopoly rents providing an incentive for firms to undertake costly innovative investments. This model has had a tremendous impact on the literature, and we could only provide a partial review of its applications. Then, we decided to focus on a few major themes: trade and biased technical change, with their effects on growth and inequality, financial development, complementarity in the process of innovation and endogenous fluctuations.

While only being a limited selection, these applications give a sense of the success of the model in providing a tractable framework for analyzing a wide array of issues in economic growth. In fact, we have shown how the model can incorporate a number of general equilibrium effects that are fundamental in the analysis of trade, wage inequality, cross-country productivity differences and other topics. Further, while the original model has linear AK-dynamics, we have surveyed recent generalizations featuring richer dynamics, which can potentially be applied to the study of financial development and innovations waves. Given its longevity, flexibility and simplicity, we are convinced that the growth model with horizontal innovation will continue to be useful in future research.

Acknowledgements

We thank Philippe Aghion, Jeremy Greenwood, Kiminori Matsuyama and Jerome Vandennbusche for helpful comments, Zheng Song for excellent research assistance, and Christina Lönnblad for editorial assistance.

Uncited references

[Jones (1999)] [Jones (2001)]

References

Acemoglu, D. (1998). "Why do new technologies complement skills? Directed technical change and wage inequality". Quarterly Journal of Economics 113, 1055–1090.
Acemoglu, D. (2002). "Directed technical change". Review of Economic Studies 69, 781–809.
Acemoglu, D. (2003a). "Patterns of skill premia". Review of Economic Studies 70, 199–230.
Acemoglu, D. (2003b). "Labor- and capital-augmenting technical change". Journal of the European Economic Association 1, 1–37.
Acemoglu, D., Zilibotti, F. (2001). "Productivity differences". Quarterly Journal of Economics 116, 563–606.

Acemoglu, D., Zilibotti, F. (1997). "Was Prometheus unbounded by chance? Risk, diversification and growth". Journal of Political Economy 105, 710–751.
Aghion, P., Howitt, P. (1998). Endogenous Growth Theory. MIT Press, Cambridge, MA.
Aghion, P., Howitt, P. (1992). "A model of growth through creative destruction". Econometrica 60 (2), 323–351.
Antras, P. (2004). "Is the U.S. aggregate production function Cobb–Douglas? New estimates of the elasticity of substitution". Contributions to Macroeconomics 4 (1).
Arrow, K.J. (1962). "The economic implications of learning-by-doing". Review of Economic Studies 29 (1), 155–173.
Atkinson, A.B., Stiglitz, J.E. (1969). "A new view of technological change". Economic Journal, 573–578.
Azariadis, C. (1993). Intertemporal Macroeconomics. Blackwell.
Azariadis, C., Guesnerie, R. (1986). "Sunspots and cycles". Review of Economic Studies 53, 725–737.
Barro, R.J., Sala-i-Martin, X. (1995). Economic Growth. McGraw-Hill.
Barro, R.J., Sala-i-Martin, X. (1997). "Technological diffusion, convergence, and growth". Journal of Economic Growth 2, 1–26.
Benassy, J.-P. (1998). "Is there always too little research in endogenous growth with expanding product variety?". European Economic Review 42, 61–69.
Bencivenga, V., Smith, B. (1991). "Financial intermediation and endogenous growth". Review of Economic Studies 58, 195–209.
Borjas, G.J., Freeman, R.B., Katz, L.F. (1997). "How much do immigration and trade affect labor market outcomes". Brookings Papers on Economic Activity, 1–67.
Braudel, F. (1979). Civilization and Capitalism. Harper and Row, New York.
Caselli, F., Esquivel, G., Lefort, F. (1996). "Reopening the convergence debate: A new look at cross-country growth empirics". Journal of Economic Growth 1, 363–389.
Caselli, F., Coleman, J. (2000), "The world technology frontier". NBER Working Paper 7904.
Cetorelli, N. (2002). "Could prometheus be bound again? A contribution to the convergence controversy". Journal of Economic Dynamics and Control 27, 29–50.
Ciccone, A., Matsuyama, K. (1999). "Efficiency and equilibrium with dynamic increasing returns due to demand complementarities". Econometrica 67, 499–525.
Ciccone, A., Matsuyama, K. (1996). "Start-up costs and pecuniary externalities as barriers to economic development". Journal of Development Economics 49, 33–59.
Cohen, W.M., Levin, R.C. (1989). "Empirical studies of innovation and market structure". In: Handbook of Industrial Organization, vol. II. Elsevier, Amsterdam, pp. 1059–1107.
Cooper, R.W., John, A. (1988). "Coordinating coordination failures in Keynesian models". Quarterly Journal of Economics 103, 441–463.
Dalgaard, C.-J., Kreiner, C.T. (2001). "Is declining productivity inevitable?". Journal of Economic Growth 6, 187–203.
Dinopoulos, E., Segerstrom, P. (1999). "A Schumpeterian model of protection and relative wages". American Economic Review 89, 450–472.
Dinopoulos, E., Segerstrom, P. (2003). "A theory of North–South trade and globalization". Mimeo. Stockholm School of Economics.
Devereux, M.B., Lapham, B.J. (1994). The stability of economic integration and endogenous growth. Quarterly Journal of Economics 109, 299–305.
DeVries, J. (1990). The Economy of Europe in an Age of Crisis; 1600–1750. Cambridge University Press, Cambridge.
Dixit, A.K., Stiglitz, J.E. (1977). "Monopolistic competition and optimum product diversity". American Economic Review 67 (3), 297–308.
Diwan, I., Rodrik, D. (1991). "Patents, appropriate technology, and North–South trade". Journal of International Economics 30, 27–48.
Dornbusch, R., Fischer, S., Samuelson, P.A. (1977). "A continuum Ricardian model of comparative advantage, trade and payments". American Economic Review 67, 823–839.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 58
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

Epifani, P., Gancia, G. (2002). "The skill bias of world trade". IIES Seminar Paper 707.
Ethier, W.J. (1982). "National and international returns to scale in the modern theory of international trade". American Economic Review 72 (3), 389–405.
Evans, G., Honkapojha, S., Romer, P. (1998). "Growth cycles". American Economic Review 88 (3), 495–515.
Evans, G., Honkapohja, S. (2001). Learning and Expectations in Macroeconomics. Princeton University Press, Princeton.
Francois, P., Lloyd-Ellis, H. (2003). "Animal spirits through creative destruction". American Economic Review 93, 530–550.
Funke, M., Strulik, H. (2000). "On endogenous growth with physical capital, human capital and product variety". European Economic Review 44, 491–515.
Gancia, G. (2003). "Globalization, divergence and stagnation". IIES Seminar Paper 720.
Grandmont, J.-M. (1986). "Periodic and aperiodic behavior in discrete one-dimensional dynamical system". In: Hildenbrand, W., Mas-Colell, A. (Eds.), Contributions to Mathematical Economics: In Honor of Gerard Debreu. Elsevier, Amsterdam.
Greenwood, J., Jovanovic, B. (1990). "Financial development, growth and the distribution of income". Journal of Political Economy 98, 1067–1107.
Griliches, Z., Schmookler, J. (1963). "Inventing and maximizing". American Economic Review 53, 725–729.
Grossman, G., Helpman, E. (1989). "Product development and international trade". Journal of Political Economy 97, 1261–1283.
Grossman, G., Helpman, E. (1991a). Innovation and Growth in the World Economy. MIT Press, Cambridge, MA.
Grossman, G., Helpman, E. (1991b). "Endogenous product cycles". Economic Journal 101, 1214–1229.
Grossman, G., Lai, E. (2004), "International protection of intellectual property", American Economic Review, in press.
Habakkuk, H.J. (1962). American and British Technology in the Nineteenth Century: Search for Labor Saving Inventions. Cambridge University Press, Cambridge.
Hall, R., Jones, C.I. (1999). "Why do some countries produce so much more output per workers than others?". Quarterly Journal of Economics 114, 83–116.
Hamermesh, D.S. (1993). Labor Demand. Princeton University Press, Princeton.
Helpman, E. (1993). "Innovation, imitation and intellectual property rights". Econometrica 61, 1247–1280.
Helpman, E. (1998). General Purpose Technology and Economic Growth. MIT Press, Cambridge, MA.
Helpman, E., Krugman, P. (1985). Market Structure and Foreign Trade. MIT Press, Cambridge, MA.
Hicks, J. (1932). The Theory of Wages. Macmillan, London.
Imbs, J., Wacziarg, R. (2003). "Stages of diversification". American Economic Review 93, 63–86.
Jones, C.I. (1995). "R&D-based models of economic growth". Journal of Political Economy 103 (August), 759–784.
Jones, C.I. (1999). "Growth: With or without scale effects". American Economic Review 89, 139–144.
Jones, C.I. (2001). "Was the industrial revolution inevitable? Economic growth over the very long run". Advances in Macroeconomics 1 (2). Article 1.
Judd, K.L. (1985). "On the performance of patents". Econometrica 53 (3), 567–585.
Kalemli-Ozcan, S., Sorensen, B.E., Yosha, O. (2001). "Economic integration, industrial specialization, and the asymmetry of macroeconomic fluctuations". Journal of International Economics 55, 107–137.
Kalemli-Ozcan, S., Sorensen, B.E., Yosha, O. (2003). "Risk sharing and industrial specialization: Regional and international evidence". American Economic Review 93, 903–918.
Kennedy, C. (1964). "Induced bias in innovation and the theory of distribution". Economic Journal 74, 541–547.
Klenow, P.J. (1996). "Industry innovation: Where and why". Carnegie–Rochester Conference Series on Public Policy 44, 125–150.
Klenow, P.J., Rodriguez-Clare, A. (1997). "The neoclassical revival in growth economics: Has it gone too far?". In: Bernanke, B.S., Rotemberg, J.J. (Eds.), NBER Macroeconomics Annual 1997. MIT Press, Cambridge, MA, pp. 73–102.

Krugman, P. (1991). "Increasing returns and economic geography". Journal of Political Economy 99, 483–499.
Krugman, P., Venables, A. (1995). "Globalization and the inequality of nations". Quarterly Journal of Economics 110, 857–880.
Krusell, P., Ohanian, L., Rios-Rull, V., Violante, G. (2000). "Capital skill complementarity and inequality". Econometrica 68, 1029–1053.
Kwan, F.Y.K., Lai, E.L.C. (2003). "Intellectual property rights protection and endogenous economic growth". Journal of Economic Dynamics and Control 27, 853–873.
Lai, E.L.C. (1998). "International intellectual property rights protection and the rate of product innovation". Journal of Development Economics 55, 133–153.
Lai, E.L.C., Qiu, L.D. (2003). "The North's intellectual property rights standard for the South?". Journal of International Economics 59, 183–209.
Levhari, D., Sheshinski, E. (1969). "A theorem on returns to scale and steady-state growth". Journal of Political Economy 77, 60–65.
Martin, P., Rey, H. (2000). "Financial integration and asset returns". European Economic Review 44, 1327–1350.
Martin, P., Rey, H. (2001). "Financial super-markets: Size matters for asset trade". NBER Working Paper 8476.
Martin, P., Rey, H. (2002). "Financial globalization and emerging markets: With or without crash?". NBER Working Paper 9288.
Matsuyama, K. (1995). "Complementarities and cumulative processes in models of monopolistic competition". Journal of Economic Literature 33, 701–729.
Matsuyama, K. (1996). "Why are there rich and poor countries? Symmetry-breaking in the world economy". Journal of the Japanese and International Economies 10, 419–439.
Matsuyama, K. (1997). "Complementarity, instability and multiplicity". The Japanese Economic Review 48, 240–266.
Matsuyama, K. (1999). "Growing through cycles". Econometrica 67, 335–348.
Matsuyama, K. (2001). "Growing through cycles in an infinitely lived agent economy". Journal of Economic Theory 100, 220–234.
Murphy, K.M., Shleifer, A., Vishny, R.W. (1989). "Industrialization and the big push". Quarterly Journal of Economics 106 (2), 503–530.
Neary, P. (2003). "Globalisation and market structure". Journal of the European Economic Association 1, 245–271.
Nordhaus, W.D. (1969a). "An economic theory of technological change". American Economic Review 59 (2), 18–28.
Nordhaus, W.D. (1969b). Invention, Growth and Welfare: A Theoretical Treatment of Technological Change. MIT Press, Cambridge, MA.
North, D., Thomas, R.P. (1973). The Rise of the Western World: A New Economic History. Cambridge University Press, Cambridge.
Parente, S.L., Prescott, E.C. (1994). "Barriers to technology adoption and development". Journal of Political Economy 102, 298–321.
Porter, M.E., Stern, S. (2000). "Measuring the "ideas" production function: Evidence from international patent output". NBER Working Paper 7891.
Prescott, E.C. (1998). "Needed: A theory of total factor productivity". International Economic Review 39, 525–553.
Ramey, G., Ramey, V. (1995). "Cross-country evidence of the link between volatility and growth". American Economic Review 85, 1138–1151.
Rebelo, S. (1991). "Long-run policy analysis and long-run growth". Journal of Political Economy 99 (3), 500–521.
Rivera-Batiz, L., Romer, P. (1991a). "Economic integration and endogenous growth". Quarterly Journal of Economics 106, 531–555.

aghion v.2005/07/15 Prn:25/07/2005; 9:04 F:aghion1003.tex; VTEX/Ramune p. 60
aid: 1003 pii: S1574-0684(05)01003-8 docsubty: REV

Rivera-Batiz, L., Romer, P. (1991b). "International trade with endogenous technological change". European Economic Review 35, 971–1004.
Rodriguez-Clare, A. (1996). "The division of labor and economic development". Journal of Development Economics 49, 3–32.
Romer, P. (1986). "Increasing returns and long-run growth". Journal of Political Economy 94, 1002–1037.
Romer, P. (1987). "Growth based on increasing returns due to specialization". American Economic Review 77, 56–62.
Romer, P. (1990). "Endogenous technological change". Journal of Political Economy 98, 71–102.
Romer, P. (1994). "New goods, old theory and the welfare costs of trade restrictions". Journal of Development Economics 43, 5–77.
Rosenberg, N. (1976). Perspectives on Technology. Cambridge University Press, Cambridge.
Samuelson, P. (1965). "A theory of induced innovations along Kennedy–Weisacker lines". Review of Economics and Statistics, 444–464.
Schmookler, J. (1966). Invention and Economic Growth. Harvard University Press.
Schumpeter, J.A. (1942). Capitalism, Socialism and Democracy. Harper, New York.
Segerstrom, P.S., Anant, T.C.A., Dinopoulos, E. (1990). "A Schumpeterian model of the product life cycle". American Economic Review 80, 1077–1091.
Shell, K. (1973). "Inventive activity, industrial organization and economic growth". In: Mirrels, J.A., Stern, N.H. (Eds.), Models of Economic Growth. Wiley, New York.
Shleifer, A. (1986). "Implementation cycles". Journal of Political Economy 94, 1163–1190.
Solow, R.M. (1956). "A contribution to the theory of economic growth". Quarterly Journal of Economics 70, 65–94.
Spence, M. (1976). "Product selection, fixed costs, and monopolistic competition". Review of Economic Studies 43, 217–235.
Stiglitz, J.E. (1970). "Factor price equalization in a dynamic economy". Journal of Political Economy 78, 456–488.
Swan, T.W. (1956). "Economic growth and capital accumulation". Economic Record 32, 334–361.
Thoenig, M., Verdier, T. (2003). "Trade induced technical bias and wage inequalities: A theory of defensive innovations". American Economic Review 93, 709–728.
Ventura, J. (1997). "Growth and interdependence". Quarterly Journal of Economics 112, 57–84.
Vernon, R. (1966). "International investment and international trade in product-cycle". Quarterly Journal of Economics 80, 190–207.
Wood, A. (1994). North–South Trade, Employment and Inequality: Changing Fortunes in a Skill Driven World. Clarendon Press, Oxford.
Yang, G., Maskus, K.E. (2001). "Intellectual property rights, licensing and innovation in an endogenous product-cycle model". Journal of International Economics 53, 169–187.
Young, A. (1928). "Increasing returns and economic progress". Economic Journal 38 (152), 527–542.
Young, A. (1991). "Learning by doing and the dynamic effects of international trade". Quarterly Journal of Economics 106, 369–406.
Young, A. (1993). "Substitution and complementarity in endogenous innovation". Quarterly Journal of Economics 108, 775–807.
Young, A. (1995). "The tyranny of numbers: Confronting the statistical realities of the East Asian growth experience". Quarterly Journal of Economics 110, 641–680.
Xie, D. (1998). "An endogenous growth model with expanding ranges of consumer goods and of producer durables". International Economic Review 39, 439–460.
Zilibotti, F. (1994). "Endogenous growth and intermediation in an archipelago economy". Economic Journal 104, 462–473.
Zilibotti, F. (1995). "A Rostovian model of endogenous growth and underdevelopment traps". European Economic Review 39, 1569–1602.

Proof of Raw Subject Index

Proof of Raw Subject Index

From Stagnation to Growth: Unified Growth Theory*

Oded Galor†

August 11, 2004

Abstract

This chapter examines the process of development from an epoch of Malthusian stagnation to a state of sustained economic growth. The analysis focuses on recently advanced unified growth theories that capture the intricate evolution of income per capita, technology, and population over the course of human history. Deciphering the underlying forces that triggered the transition from stagnation to growth and the associated phenomenon of the great divergence in income per capita across countries has been widely viewed as one of the most significant challenges facing researchers in the field of growth and development. The inconsistency of non-unified growth models with the main characteristics of the process of development across most of human history induced growth theorists to advance an alternative theory that captures in a single unified framework the epoch of Malthusian stagnation, the modern era of sustained economic growth, and the recent transition between these distinct regimes. Unified growth theory reveals the underlying micro foundations that are consistent with the growth process over the entire history of the human species, enhancing the confidence in the viability of the theory, its predictions and policy implications for the growth process of less developed economies.

Keywords: Growth, Technological Progress, Demographic Transition, Income Distribution, Human Capital, Evolution, Natural Selection, Malthusian Stagnation, Class Structure.

JEL classification Numbers: O11, O14, O33, O40, J11, J13.

---

*Forthcoming in the *Handbook of Economic Growth* (P. Aghion and S. Durlauf eds.), North-Holland, 2004.

†The author wishes to thank Philippe Aghion, Graziella Bertocchi, Carl Johan Dalgaard, Matthias Doepke, Hagai Etkes, Moshe Hazan, Nils-Petter Lagerloef, Sebnem Kalemli-Ozcan, Daniel Mejia, Joel Mokyr, Omer Moav, Andrew Mountford, Nathan Sussman, and David Weil for valuable discussions and detailed comments, and Tamar Roth for excellent research assistance. This research is supported by *NSF Grant SES-0004304.*

Table of Content

"A complete, consistent, unified theory...would be the ultimate triumph of human reason"
Stephen W. Hawking - A Brief History of Time.

1 Introduction

This chapter examines the recent advance of a *unified growth theory* that is designed to capture the process of development over the entire course of human history.

The inconsistency of exogenous and endogenous neoclassical growth models with some of the most fundamental features of process of development, has led recently to a search for a unified theory that would unveil the underlying micro-foundations of the growth process in its entirety, capturing the epoch of Malthusian stagnation that characterized most of human history, the contemporary era of modern economic growth, and the underlying driving forces that triggered the recent transition between these regimes and the associated phenomenon of the Great Divergence in income per capita across countries.

The evolution of economies over the major portion of human history was marked by Malthusian Stagnation. Technological progress and population growth were miniscule by modern standards and the average growth rate of income per capita in various regions of the world was even slower due to the offsetting effect of population growth on the expansion of resources per capita. In the past two centuries, in contrast, the pace of technological progress increased significantly in association with the process of industrialization. Various regions of the world economy departed from the Malthusian trap and experienced initially a considerable rise in the growth rates of income per capita and population. Unlike episodes of technological progress in the pre-Industrial Revolution era that failed to generate sustained economic growth, the increasing role of human capital in the production process in the second phase of the Industrial Revolution ultimately prompted a demographic transition, liberating the gains in productivity from the counterbalancing effects of population growth. The decline in population growth and the associated enhancement in technological progress and human capital formation paved the way for the emergence of the modern state of sustained economic growth.

The fundamental factors that generated the remarkable escape from the Malthusian trap have been shrouded in mystery until recently and their significance for the understanding of the contemporary growth process have been under-explored.

- What accounts for the epoch of stagnation that characterized most of human history?
- What is the origin of the sudden spurt in growth rates of output per capita and population?
- Why had episodes of technological progress in the pre-industrialization era failed to generate sustained economic growth?
- What was the source of the dramatic reversal in the positive relationship between income per capita and population that existed throughout most of human history?
- What triggered the demographic transition?
- Would the transition to a state of sustained economic growth have been feasible without the demographic transition?
- What are the underlying behavioral and technological structures that can simultaneously account for these distinct phases of development and what are their implications for the contemporary growth process of developed and underdeveloped countries?

The perplexing phenomenon of the Great Divergence in income per capita across regions of the world in the past two centuries presents an additional mystery about the growth process.

- What accounts for the sudden take-off from stagnation to growth in some countries in the world and the persistent stagnation in others?
- Why has the positive link between income per capita and population growth reversed its course in some economies but not in others?
- Why have the differences in per capita incomes across countries increased so markedly in the last two centuries?
- Has the transition to a state of sustained economic growth in advanced economies adversely affected the process of development in less-developed economies?

Deciphering the fundamental determinants of the transition from stagnation to growth and the great divergence has been widely viewed as one of the most significant research challenges facing researchers in the field of growth and development.

The transitions from a Malthusian epoch to a state of sustained economic growth and the related phenomenon of the Great Divergence, as depicted in Figure 2.1, have significantly shaped the contemporary world economy.[1] Nevertheless, the distinct qualitative aspects of the growth process during most of human history were virtually ignored in the shaping of neoclassical growth models, resulting in a growth theory that is only consistent with a small fragment of human history.

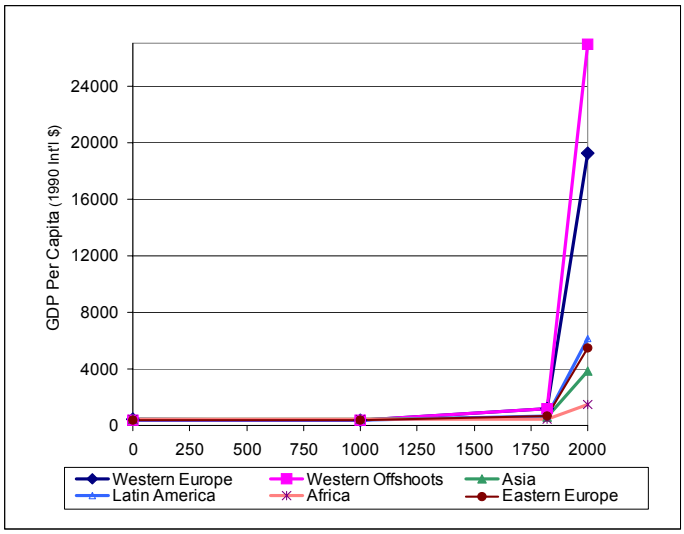


Figure 2.1. The Evolution of Regional Income per Capita over the Years 1 - 2001

Sources: Maddison (2003)[2]

[1]The ratio of GDP per capita between the richest region and the poorest region in the world was only 1.1:1 in the year 1000, a 2:1 in the year 1500 and 3:1 in the year 1820. In the course of the 'Great Divergence' the ratio of GDP per capita between the richest region (Western offshoots) and the poorest region (Africa) has widened considerably from a modest 3:1 ratio in 1820, to a 5:1 ratio in 1870, a 9:1 ratio in 1913, a 15:1 in 1950, and a huge 18:1 ratio in 2001.

[2]According to Maddison's classification, "Western Offshoots" consists of United States, Canada, Australia and New Zealand.

The preoccupation of growth theory with the empirical regularities that have characterized the growth process of developed economies in the past century and of less developed economies in the last few decades, has become harder to justify from a scientific viewpoint in light of the existence of vast evidence about qualitatively different empirical regularities that characterized the growth process over most of human existence. Is there a scientific justification for the use of selective observations, only about the contemporary growth process, in formulating theory about the current growth process? Could we be confident about the predictions of a theory that is not based on micro foundations that are consistent with the main elements of the entire growth process? As argued by Copernicus, *"it is as thuoght an artist were to gather the hands, feet, head and other members for his images from diverse models, each part perfectly drawn, buut not related to a single body, and since they in no way match each other, the result would be monster rather than man"*.[3] The evolution of theories in older scientific disciplines suggests that theories that are founded on the basis of a subset of the existing observations and their driving forces, may be attractive in the short run, but non-robust and ultimately non-durable in the long run.[4] The attempts to develop unified theories in physics have been based on the conviction that all physical phenomena should ultimately be explainable by some underlying unity.[5] Similarly, the entire process of development and its basic causes ought to be captured by a unified growth theory.

In recent years, it has been increasingly recognized that the understanding of the contemporary growth process would be fragile and incomplete unless growth theory could be based on proper micro-foundations that reflect the qualitative aspects of the growth process and their central driving forces. Moreover, it has become apparent that a comprehensive understanding of the hurdles faced by less developed economies in reaching a state of sustained economic growth would be futile unless the factors that prompted the transition of the currently developed economies into a state of sustained economic growth could be identified and their implications would be modified to account for the differences in the growth structure of less developed economies in an interdependent world.

The transition from stagnation to growth and the associated phenomenon of the great divergence have been the subject of an intensive research in the growth literature in recent years.[6] The inconsistency of exogenous and endogenous neoclassical growth models with the process of development along most of human history, induced growth theorists to search for an alternative theory that could capture in a single unified framework the contemporary era of sustained economic growth, the epoch of Malthusian stagnation that had characterized most of human history, and the driving forces that brought about the recent transition between these distinct regimes.

Imposing the constraint that a single theory should account for the entire intricate process of development and its prime causes in the last thousands of years is a discipline that would enhance the viability of growth theory. A unified theory of economic growth reveals the fundamental micro-foundations that are consistent with the process of economic development over the entire course of human history, rather that with the last century only, boosting the confidence in growth theory, its predictions and policy implications. Moreover, it improves the understanding of the underlying factors

[3] Quoted in Kuhn (1957).

[4] For instance, Classical Thermodynamics that lacked micro-foundations was ultimately superseded by the micro-based Statistical Mechanics.

[5] *Unified Field Theory*, for instance, proposes to unify by a set of general laws the four distinct forces that are known to control all the observed interactions in matter: electromagnetism, gravitation, the weak force, and the strong force. The term unified field theory was coined by Einstein, whose research on relativity had led him to the hypothesis that it should be possible to find a unifying theory for the electromagnetic and gravitational forces.

[6] The transition from Malthusian stagnation to sustained economic growth was explored by Galor and Weil (1999, 2000), Lucas (2002), Galor and Moav (2002), Hansen and Prescott (2002), Jones (2001), Stokey (2002), as well as others, and the association of Great Divergence with the transition was analyzed by Galor and Mountford (2003).

that led to the transition from stagnation to growth of the currently developed countries, shedding light on the growth process of the less developed economies.

2 Historical Evidence

This section examines the historical evidence about the evolution of the relationship between income per capita, population growth, technological change and human capital formation along three distinct regimes that have characterized the process of economic development: The Malthusian Epoch, The Post-Malthusian Regime, and the Sustained Growth Regime.

During the Malthusian Epoch that characterized most of human history, technological progress and population growth were insignificant by modern standards. The level of income per capita had a positive effect on population and the average growth rate of income per capita in the long-run, as depicted in Figure 2.2, was negligible due to the slow pace of technological progress as well as the counterbalancing effect of population growth on the expansion of resources per capita. During the Post Malthusian Regime, the pace of technological progress markedly increased in association with the process of industrialization, triggering a take-off from the Malthusian trap. The growth rate of income per capita increased significantly, as depicted in Figures 2.1 and 2.2, but the positive Malthusian effect of income per capita on population growth was still maintained, generating a sizeable increase in population growth that offset some of the potential gains in income per capita. The acceleration in the rate of technological progress in the second phase of the Industrial Revolution, and its interaction with human capital formation ultimately prompted the demographic transition. The rise in aggregate income was not counterbalanced by population growth, enabling technological progress to bring about sustained increase in income per capita.

2.1 The Malthusian Epoch

For thousand of years, humans were subjected to persistent struggle for existence. Survival, argued Malthus (1798), necessitated a *"perpetual struggle for room and food."* Resources generated by technological progress and land expansion were channeled primarily towards an increase in the size of the population, with a minor long-run effect on income per capita. Thus, as reflected in the viewpoint of a prominent observer of the period, *"the most decisive mark of the prosperity of any country is the increase in the number of its inhabitants"* (Smith 1776).

The evolution of population and output per capita across most of human history was consistent with the Malthusian paradigm. The positive effect of the standard of living on population growth along with diminishing labor productivity kept income per capita in the proximity of a subsistence level.[7] Periods marked by the absence of changes in the level of technology or in the availability of land, were characterized by a stable population size as well as a constant income per capita, whereas periods characterized by improvements in the technological environment or in the availability of land generated temporary gains in income per capita, leading ultimately to a larger but not richer population. Technologically superior countries had eventually denser populations but their standard of living did not reflect the degree of their technological advancement.

[7]This subsistence level of consumption may be well above the minimal physiological requirements that are necessary in order to sustain an active human being.

2.1.1 Income per capita

During the Malthusian epoch the average growth rate of output per capita was negligible and the standard of living did not differ greatly across countries. As depicted in Figure 2.2 the average level of income per capita during the first millennium fluctuated around $450 per year, and the average growth rate of output per capita in the world was nearly zero.[8] This state of Malthusian stagnation persisted until the end of the 18th century. The average level of income per capita in the world economy remained below $670 per year in the years 1000-1820 and the average growth rate of the world income per capita was miniscule, creeping at a rate of about 0.05% per year (Maddison 2001).

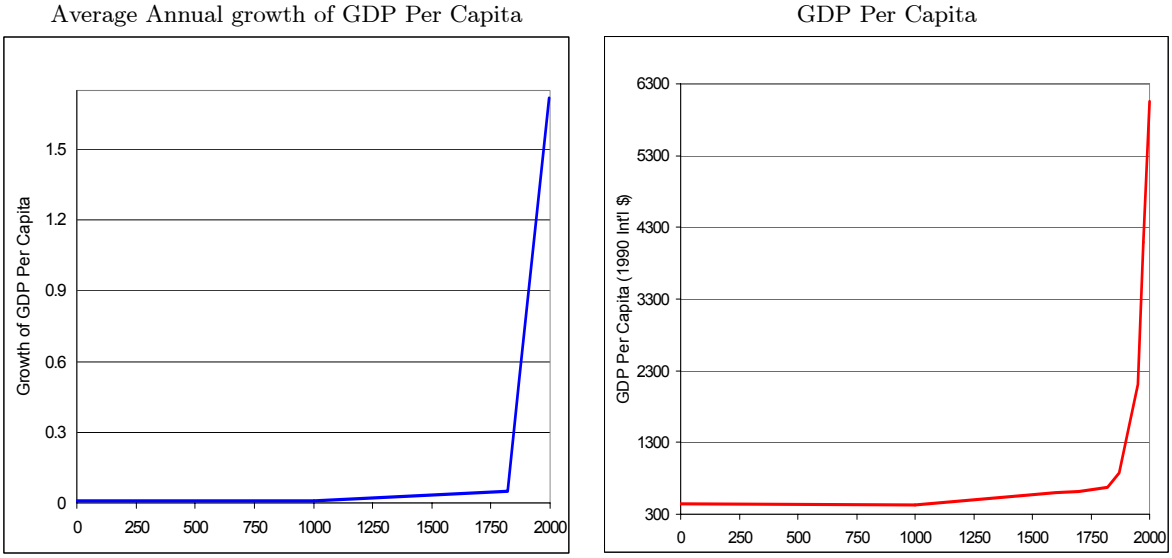


Figure 2.2. The Evolution of the World Income Per Capita over the Years 1-2001
Source: Maddison (2001, 2003)

This pattern of stagnation was observed across all regions of the world. As depicted in Figure 2.1, the average level of income per capita in Western and Eastern Europe, the Western Offshoots, Asia, Africa, and Latin America was in the range of $400-450 per year in the first millennium and the average growth rates in each of these regions was nearly zero. This state of stagnation persisted until the end of the 18th century across all regions and the average level of income per capita in the years 1000-1820 ranged from $418 per year in Africa, $581 in Asia, $692 in Latin America, and $683 in Eastern Europe, to $1202 in the Western Offshoots, and $1204 in Western Europe. Furthermore, the average growth rate of output per capita over this period ranged from 0% in the impoverish region of Africa to a sluggish rate of 0.14% in the prosperous region of Western Europe.

Despite the stability in the evolution of the world income per capita in the Malthusian epoch, from a perspective of a millennium, wages and income per capita had fluctuated significantly within regions deviating from their sluggish long-run trend over decades and sometimes over few centuries. In particular, as depicted in Figure 2.3, real GDP per capita in England fluctuated drastically over most of the past millennium. It declined during the 13th century, and increased sharply during the 14th and the 15th century in response to the catastrophic population decline in the aftermath of the Black Death. This two-century rise in per capita real income stimulated population growth and brought about a decline in income per capita in the 16th century back to its level in the first half of the 14th century.

[8]Maddison's estimates of income per capita are evaluated in terms of 1990 international dollars.

Real income per capita increased once again in the 17th century and remained stable during the 18th century, prior to its rise during the take-off from the Malthusian epoch in the 19th century.

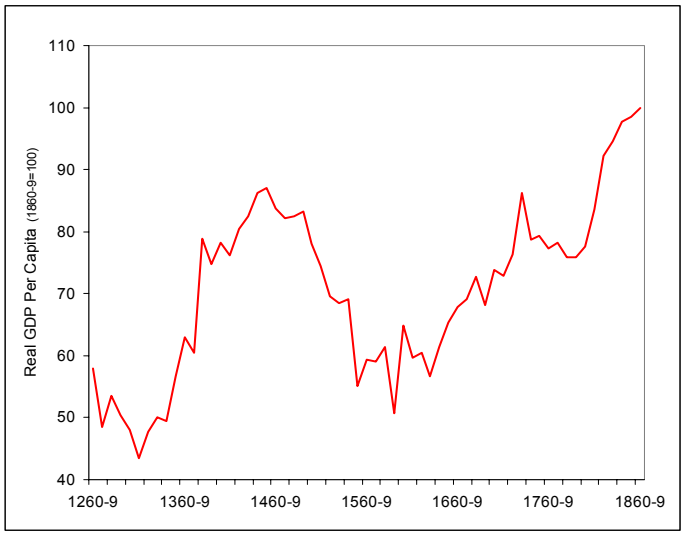


Figure 2.3. Fluctuations in Real GDP Per Capita: England, 1260-1870
Source: Clark (2001)

2.1.2 Income and Population Growth

Population Growth and Income level

Population growth over this Malthusian era followed the Malthusian pattern as well. As depicted in Figures 2.4 and 2.5, the slow pace of resource expansion in the first millennium was reflected in a modest increase in the population of the world from 231 million people in the year 1 to 268 million in the year 1000; a miniscule average growth rate of 0.02% per year.[9] The more rapid (but still very slow) expansion of resources in the period 1000-1500, permitted the world population to increase by 63% over this period, from 268 million in the year 1000 to 438 million in the year 1500; a slow 0.1% average growth rate per year. Resource expansion over the period 1500-1820 had a more significant impact on the world population, that grew 138% from 438 million in the year 1500 to 1041 million in the year 1820; an average pace of 0.27% per year.[10] This positive effect of income per capita on the size of the population was maintained in the last two centuries as well, as world population reached a remarkable level of nearly 6 billion people.

[9]Since output per capita grew at an average rate of 0% per year over the period 0-1000, the pace of resource expansion was approximately equal to the pace of population growth, namely, 0.02% per year.

[10]Since output per capita in the world grew at an average rate of 0.05% per year in the time period 1000-1500 as well as in the period 1500-1820, the pace of resource expansion was approximately equal to the sum of the pace of population growth and the growth of output per capita. Namely, 0.15% per year in the period, 1000-1500 and 0.32% per year in the period 1500-1820.

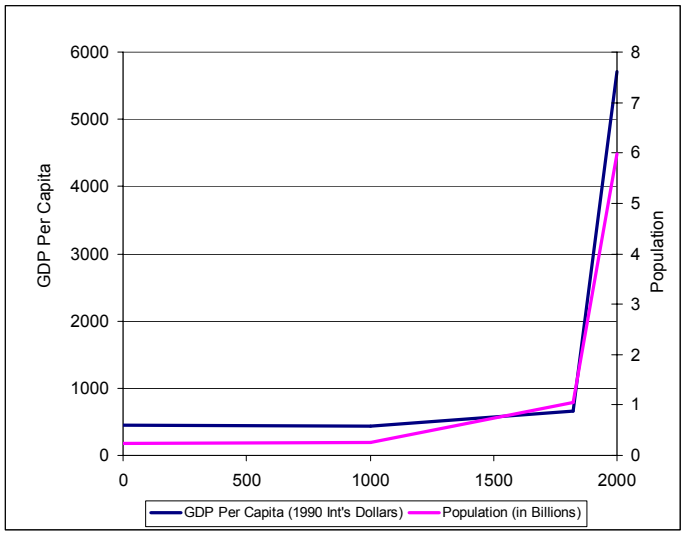


Figure 2.4. The Evolution of World Population and Income Per Capita over the Years 1 - 2000
Source: Maddison (2001)

Moreover, the gradual increase in income per capita during the Malthusian epoch was associated with a monotonic increase in the average rate of growth of world population, as depicted in Figure 2.5.[11]

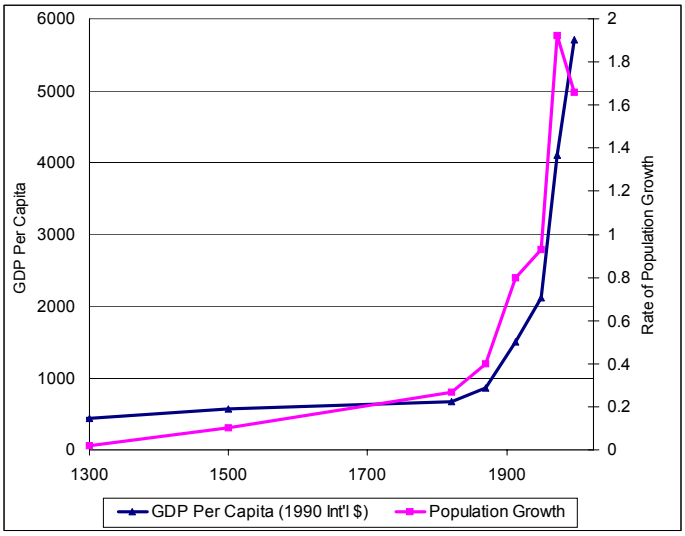


Figure 2.5. Population Growth and Income Per Capita in the World Economy
Source: Maddison (2001)

Fluctuations in Income and Population

Fluctuations in population and wages exhibited the Malthusian pattern as well. Episodes of technological progress, land expansion, favorable climatic conditions, or major epidemics (that resulted in a decline of the adult population), brought about a temporary increase in real wages and income per

[11]Lee (1997) reports positive income elasticity of fertility and negative income elasticity of mortality from studies examining a wide range of pre-industrial countries. Similarly, Wrigley and Schofield (1981) find a strong positive correlation between real wages and marriage rates in England over the period 1551-1801. Clark (2003) finds that in England, at the beginning of the 17th century, the number of surviving offspring is higher among households with higher level of income and literacy rates, suggesting that the positive effect of income on fertility is present cross-sectionally, as well.

capita. In particular, as depicted in Figure 2.6, the catastrophic decline in the population of England during the Black Death (1348-1349), from about 6 millions to about 3.5 millions people, increased significantly the land-labor ratio, tripling real wages in the subsequent 150 years. Ultimately, however, most of this increase in real resources per capita was channelled towards increased fertility rates, increasing the size of the population, and bringing the real wage rate in the 1560s back to the proximity of its pre-plague level.[12]

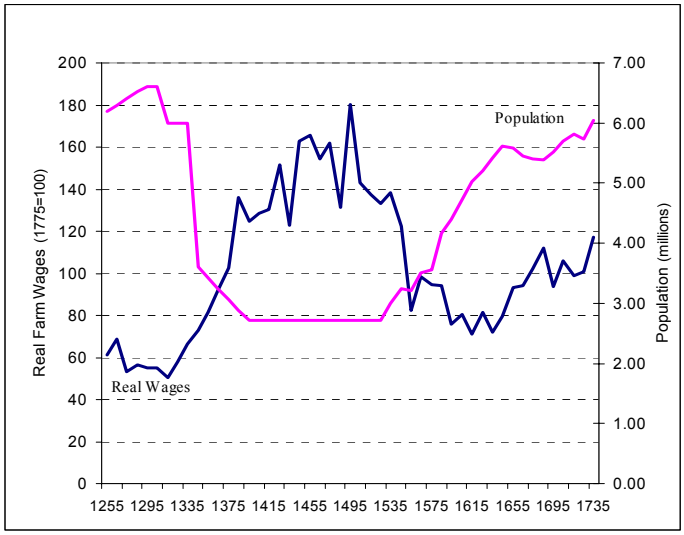


Figure 2.6. Population and Real Wages: England, 1250-1750
Source: Clark (2001, 2002)

Population Density

Variations in population density across countries during the Malthusian epoch reflected primarily cross country differences in technologies and land productivity. Due to the positive adjustment of population to an increase in income per capita, differences in technologies or in land productivity across countries resulted in variations in population density rather than in the standard of living.[13] For instance, China's technological advancement in the period 1500-1820 permitted its share of world population to increase from 23.5% to 36.6%, while its income per capita in the beginning and the end of this time interval remained constant at roughly $600 per year.[14]

This pattern of increased population density persisted until the demographic transition, namely, as long as the positive relationship between income per capita and population growth was maintained. In the period 1600-1870, United Kingdom's technological advancement relative to the rest of the world more than doubled its share of world population from 1.1% to 2.5%. Similarly, in the period 1820-1870, the land abundant, technologically advanced, economy of the US. experienced a 220% increased in its share of world population from 1% to 3.2%.[15]

[12]Reliable population data is not available for the period 1405-1525 and figure 2.6 is depicted under the assumption maintained by Clark (2001) that population was rather stable over this period.

[13]Consistent with the Malthusian paradigm, China's sophisticated agricultural technologies, for example, allowed high per-acre yields, but failed to raise the standard of living above subsistence. Similarly, the introduction of the potato in Ireland in the middle of the 17th century generated a large increase in population over two centuries without a significant improvements in the standard of living. Furthermore, the destruction of potatoes by fungus in the middle of the 19th century, generated a massive decline in population due to the Great Famine and mass migration [Mokyr 1985].

[14]The Chinese population more than tripled over this period, increasing from 103 million in the year 1500 to 381 million in the year 1820.

[15]The population of the United Kingdom nearly quadrupled over the period 1700-1870, increasing from 8.6 million in

Mortality and Fertility

The Malthusian demographic regime was characterized by fluctuations in fertility rates, reflecting variability in income per capita as well as changes in mortality rates. Periods of rising income per capita permitted a raise the number of surviving offspring, inducing an increase in fertility rates along with a reduction in mortality rates, due to improved nourishment, and health infrastructure. Periods of rising mortality rates (e.g., the black death) induced an increase in fertility rates so as to maintain the number of surviving offspring that can be supported by existing resources.

The relationship between fertility and mortality during the Malthusian epoch was complex. Demographic patterns in England during the 14th and 15th centuries, as depicted in Figure 2.6, suggest that an (exogenous) increase in mortality rates was associated with a significant rise in fertility rates. However, the period 1540-1820 in England, vividly demonstrates a negative relationship between mortality rates and fertility rates. As depicted in Figure 2.7, an increase in mortality rates over the period 1560-1650 was associated with a decline in fertility rates, whereas a decline in mortality rates in the time period 1680-1820 was associated with increasing fertility rates.

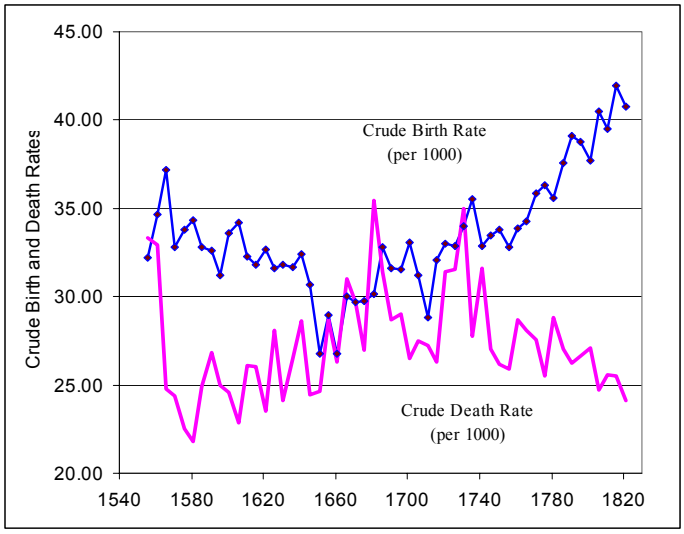


Figure 2.7. Fertility and Mortality: England 1540-1870
Source: Wrigley and Schofield (1981)

Life Expectancy

Life expectancy at birth fluctuated in the Malthusian epoch, ranging from 24 in Egypt in the time period 33 - 258 AD, to 42 in England at the end 16th century. In the initial process of European urbanization, the percentage of urban population increased six-fold from about 3% in 1520 to nearly 18% in 1750 (de Vries (1984) and Bairoch (1988)). This rapid increase in population density, without significant changes in health infrastructure, generated a rise in mortality rates and a decline in life expectancy. As depicted in Figures 2.7 and 2.8, over the period 1580-1740 mortality rates increased by 50% from 0.022 to 0.032, and life expectancy at birth fell from 42 to 28 (Wrigley and Schofield, 1981). A decline in mortality along with a rise in life expectancy began in the 1740s. Life expectancy at birth rose from 28 to 41 in England and from 25 to 40 in France over the period 1740-1830 (Livi-Bacci 1997).

the year 1700 to 31.4 million in the year 1870. Similarly, the population of the United states increased 40-fold, from 1.0 million in the year 1700 to 40.2 million in the year 1870, due to a significant labor migration, as well as high fertility rates.

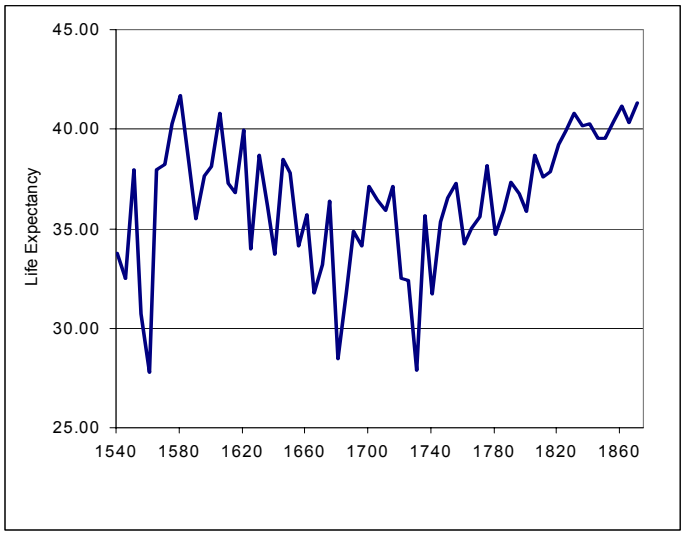


Figure 2.8. Life Expectancy: England, 1540-1870
Source: Wrigley and Schofield (1981)

2.2 The Post-Malthusian Regime

The pace of technological progress markedly increased along with the process of industrialization, instigating a take-off from the Malthusian epoch.[16] The growth rate of output per capita increased significantly, as depicted in Figures 2.1-2.3, but the positive Malthusian effect of income per capita on population growth was still maintained, generating a sizeable increase in population growth, as depicted in Figure 2.4 and 2.5, offsetting some of the gains in income.

The take-off of developed regions from the Malthusian regime was associated with the Industrial Revolution and occurred in the beginning of the 19th century, whereas the take-off of less developed regions occurred towards the beginning of the 20th century and was delayed in some countries well into the 20th century. The Post-Malthusian Regime ended with the decline in population growth in Western Europe and the Western Offshoots towards the end of the 19th century and in less developed regions in the second half of the 20th century.

2.2.1 Income Per Capita

During the Post-Malthusian Regime the average growth rate of output per capita increased significantly and the standard of living started to differ considerably across countries. As depicted in Figure 2.2, the average growth rate of output per capita in the world increased from 0.05% per year in the time period 1500-1820 to 0.53% per year in 1820-1870, and 1.3% per year in 1870-1913. The timing of the take-off and its magnitude differed across regions. As depicted in Figure 2.9, the take-off from the Malthusian Epoch and the transition to the Post-Malthusian Regime occurred in Western Europe, the Western Offshoots, and Eastern Europe at the beginning of the 19th century, whereas in Latin America, Asia (excluding China) and Africa it took place at the end of the 19th century.

[16] Ironically, it was only shortly before the time that Malthus wrote, that some regions in the world began to emerge from the trap that he described.

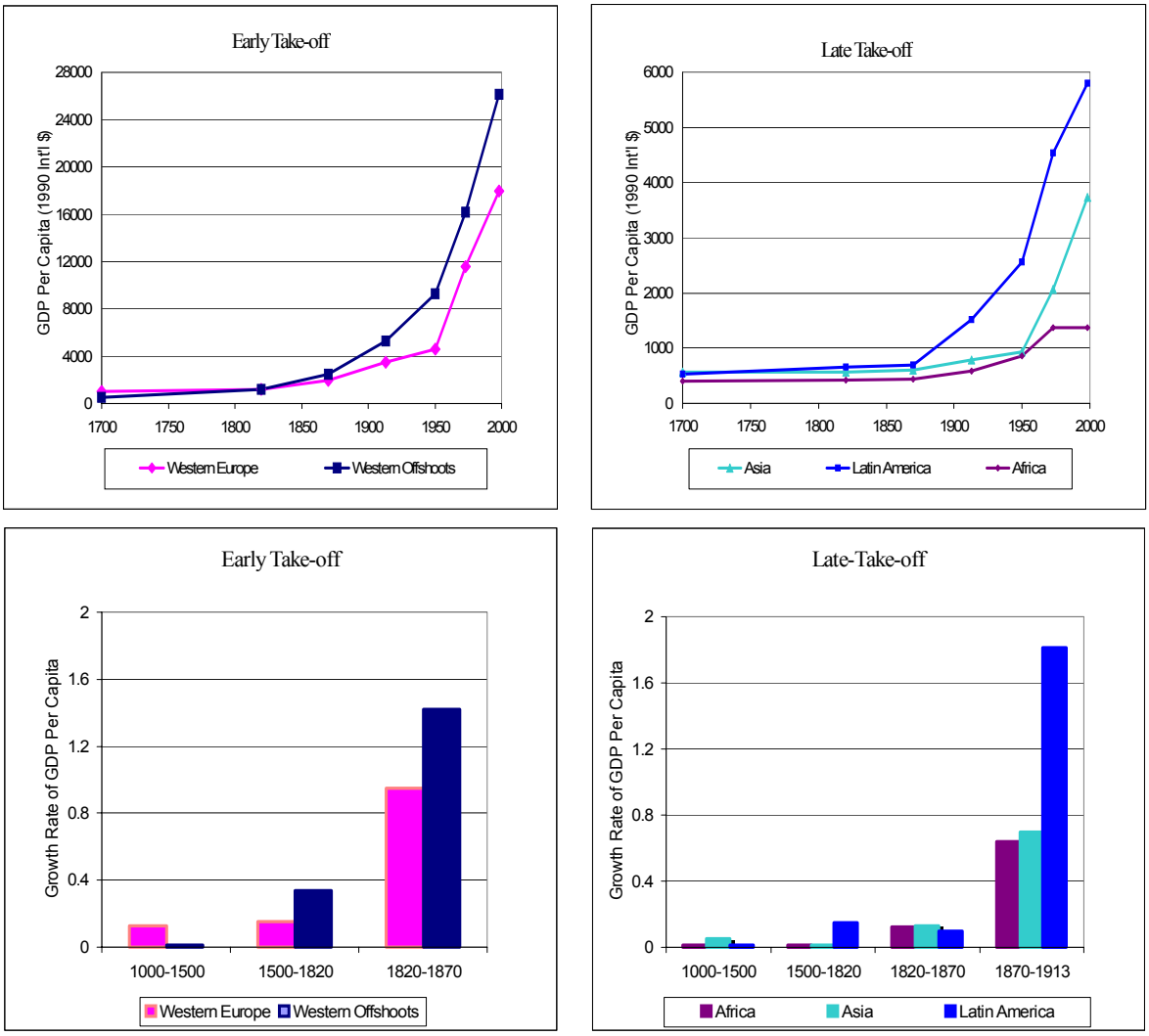


Figure 2.9. The Differential Timing of the Take-off Across Regions.
Source: Maddison (2001)

Among the regions that took off at the beginning of the 19th century, the growth rate of income per capita in Western Europe increased from 0.15% per year in the years 1500-1820 to 0.95% per year in the time period 1820-1870, and the growth rates of income per capita of the Western Offshoots increased over this period from 0.34% per year to 1.42% per year. In contrast, the take-off in Eastern Europe was more modest, and its growth rate increased from 0.1% per year in the period 1500-1820 to 0.63% per year in the time interval 1820-1870. Among the regions that took-off towards the end of the 19th century, the average growth rate of income per capita in Latin America jumped from a sluggish rate of 0.11% per year in the years 1820-1870 to a considerable 1.81% per year in the time period 1870-1913, whereas Africa's growth rates increased more modestly from 0.12% per year in the years 1820-1870 to 0.64% per year in time interval 1870-1913 and 1.02% per year in the period 1913-1950. Asia's (excluding Japan, China and India) take-off was modest as well, increasing from 0.13% per year in the years 1820-1870 to 0.64% per year in the 1870-1913 period.[17]

[17]Japan's average growth rate increased from 0.19% per year in the period 1820-1870, to 1.48% per year in the period

The level of income per capita in the various regions of the world, as depicted in Figure 2.1, ranged in the year 1870 from $444 in Africa, $543 in Asia, $698 in Latin America, and $871 in Eastern Europe, to $1974 in Western Europe and $2431 in the Western Offshoots. Thus, the differential timing of the take-off from the Malthusian epoch, increased the gap between the richest regions of Western Europe and the Western Offshoots to the impoverished region of Africa from about 3:1 in 1820 to approximately 5:1 in 1870.

The acceleration in technological progress and the accumulation of physical capital and to a lesser extent human capital, generated a gradual rise in real wages in the urban sector and (partly due to labor mobility) in the rural sector as well. As depicted in Figure 2.10, the take-off from the Malthusian epoch in the aftermath of the Industrial Revolution was associated in England with a modest rise in real wages in the first decades of the 19th century and a very significant rise in real wages after 1870.[18] A very significant rise in real wages was experienced by France, as well, after 1860.

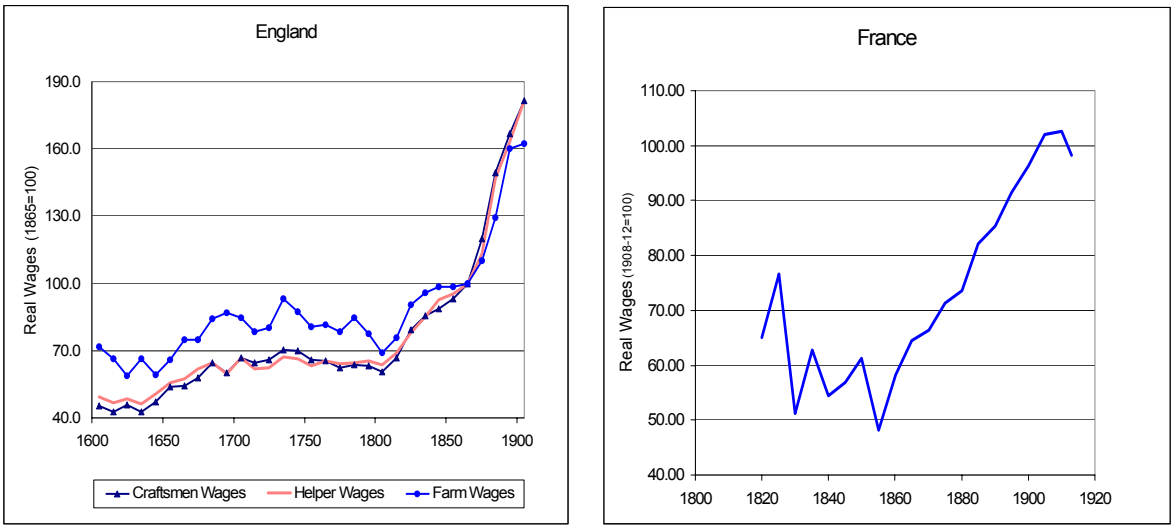


Figure 2.10. Real Wages in England and France During the take-off from the Malthusian Epoch
Sources: Clark (2002) for England, and Levy-Leboyer and Bourguignon (1990) for France

2.2.2 Income and Population Growth

The rapid increase in income per capita in the Post-Malthusian Regime was channeled partly towards an increase in the size of the population. During this Regime, the Malthusian mechanism linking higher income to higher population growth continued to function, but the effect of higher population on diluting resources per capita, and thus lowering income per capita, was counteracted by the acceleration in technological progress and capital accumulation, allowing income per capita to rise despite the offsetting effects of population growth.

The Western European take-off along with that of the Western Offshoots brought about a sharp increase in population growth in these regions and consequently a modest rise in population growth in the world as a whole. The subsequent take-off of less developed regions and the associated increase in their population growth brought about a significant rise in the population growth in the world. The

1870-1913. India's growth rate increased from 0% per year to 0.54% per year over this period , whereas China's take-off was delayed till the 1950s.

[18] Stokey (2001)'s quantitative study attributes about half of the rise in real wage over the period 1780-1850 to the forces of international trade. Moreover, technological change in manufacturing was 3 times as important as technological change in the energy sector in contributing to output growth.

rate of population growth in the world increased from an average rate of 0.27% per year in the period 1500-1820 to 0.4% per year in the years 1820-1870, and to 0.8% per year in the time interval 1870-1913. Furthermore, despite the decline in population growth in Western Europe and the Western Offshoots towards the end of the 19th century and the beginning of the 20th century, the delayed take-off of less developed regions and the significant increase in their income per capita prior to their demographic transitions generated a further increase in the rate of population growth in the world to 0.93% per year in the years 1913-1950 and a sharp rise to a high rate 1.92% per year in the period 1950-1973. Ultimately, the onset of the demographic transition in less developed economies in the second half of the 20th century, gradually reduced population growth rates to 1.66% per year in the 1973-1998 period (Maddison 2001).

Growth in Income Per Capita and Population Growth

As depicted in Figure 2.11, the take-off in the growth rate of income per capita in all regions of the world was associated with a take-off in population growth. In particular, the average growth rates of income per capita in Western Europe over the time period 1820-1870 rose to an annual rate of 0.95% (from 0.15% in the earlier period) along with a significant increase in population growth to an annual rate of about 0.7% (from 0.26% in the earlier period). Similarly, the average growth rates of income per capita in the Western Offshoots over the years 1820-1870 rose to an annual rate of 1.42% (from 0.34% in the earlier period) along with a significant increase in population growth to an annual rate of about 2.87% (from 0.43% in the earlier period).

A similar pattern is observed in Asia, and as depicted in Figure 2.11 in Africa and Latin America as well. The average growth rates of income per capita in Latin America over the years 1870-1913 rose to an annual rate of 1.81% (from 0.1% in the period 1820-1870) and subsequently by 1.43% in time interval 1913-1950 and 2.52% in the time period 1950-1973 along with a significant increase in population growth to an annual rate of 1.64% in the period 1870-1913, 1.97% in the years 1913-1950, and 2.73% in the period 1950-1973, prior to the decline in the context of the demographic transition. Similarly, the average growth rates of income per capita in Africa over the 1870-1913 period rose to an annual rate of 0.64%, (from 0.12% in the period 1820-1870) and subsequently by 1.02% in the years 1913-1950 and 2.07% in the period 1950-1973 along with a monotonic increase in population growth from a modest average annual rate of 0.4% in the years 1820-1870, to a 0.75% in the years 1870-1913, 1.65% in the years 1913-1950, 2.33% in the time interval 1950-1973, and a rapid average annual rate of 2.73% in the 1973-1998 period .

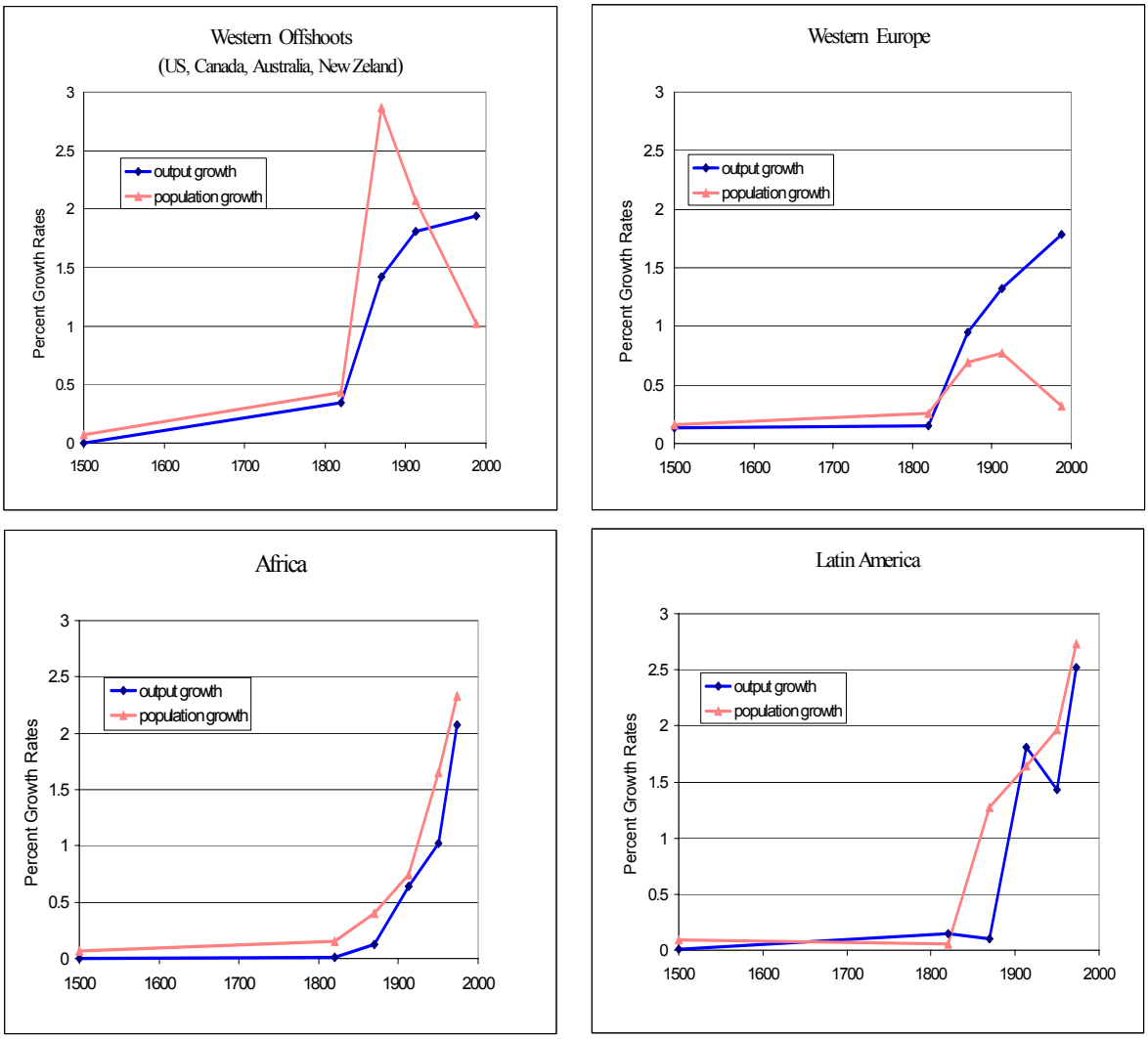


Figure 2.11. Regional Growth of GDP Per Capita and Population: 1500-2000
Source: Maddison (2001)

Ultimately, however, most regions experienced a demographic transition and thereby a transition to a state of sustained economic growth that enabled economies to convert a larger share of the fruits of factor accumulation and technological progress into the growth of output per capita.

Technological leaders and land-abundant regions during the Post-Malthusian era improved their relative position in the world in terms of their level of income per capita as well as their population size. The increase in population density of technological leaders persisted as long as the positive relationship between income per capita and population growth was maintained. Western Europe's technological advancement relative to the rest of the world increased its share of world population by 16% from 12.8% in 1820 to 14.8% in 1870, where the regional technological leader, the United Kingdom, increased its share of world population by 25% (from 2% to 2.5%) over this fifty year period. Moreover, land abundance and technological advancement in the Western Offshoots (US, Australia, New Zealand and Canada) increased their share of world population by 227% over a fifty year period, from 1.1% in 1820 to 3.6% in 1870.

The rate of population growth relative to the growth rate of aggregate income declined gradually over the period. For instance, the growth rate of total output in Europe was 0.3% per year between 1500

and 1700, and 0.6% per year between 1700 and 1820. In both periods, two thirds of the increase in total output was matched by increased population growth, and the growth of income per capita was only 0.1% per year in the earlier period and 0.2% in the later one. In the United Kingdom, where growth was the fastest, the same rough division between total output growth and population growth can be observed: total output grew at an annual rate of 1.1% in the 120 years after 1700, while population grew at an annual rate of 0.7%. Population and income per capita continued to grow after 1820, but increasingly the growth of total output was expressed as growth of income per capita. Population growth was 40% as large as total output growth over the time period 1820-1870, dropping further after the demographic transition to about 20% of output growth over the 1929-1990 period.

Fertility and Mortality

The relaxation in the households' budget constraints in the Post-Malthusian Regime permitted an increase in fertility rates along with an increase in literacy rates and years of schooling. Despite the decline in mortality rates, fertility rates (as well as population growth) increased in most of Western Europe until the second half of the 19th century (Coale and Treadway (1986)).[19] In particular, as depicted in Figure 2.12, in spite of a century of a decline in mortality rates, the crude birth rates in England increased over the 18th century and the beginning of the 19th century. Thus, the *Net Reproduction Rate (*i.e., the number of daughters per woman who reach the reproduction age) increased for about the replacement level of 1 surviving daughters per women in 1740 to about 1.5 surviving daughters per woman in the eve of the demographic transition in 1870.

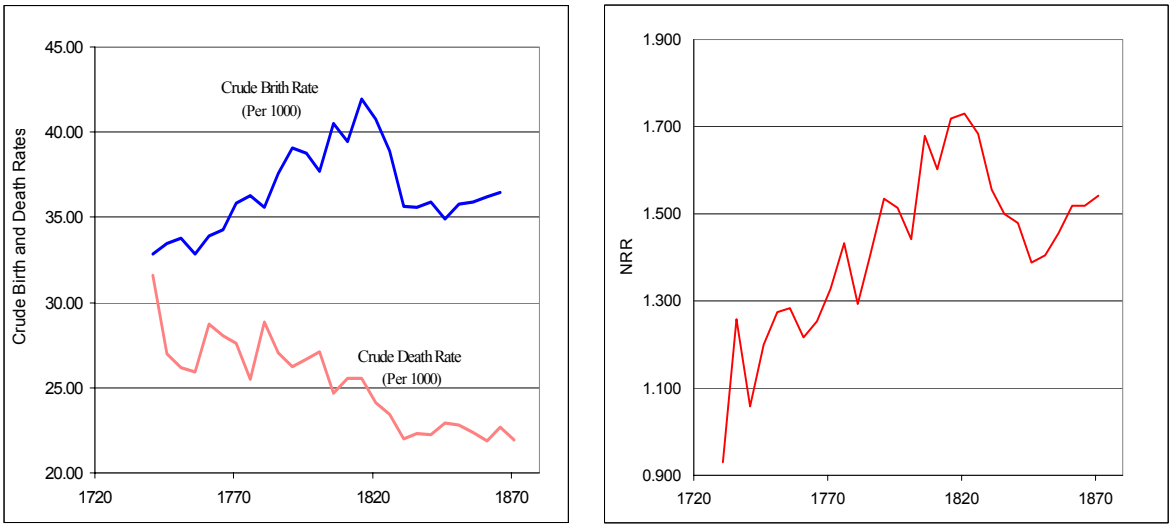


Figure 2.12. Fertility, Mortality and Net Reproduction Rate: England, 1730-1871
Source: Wrigley and Schofield (1981)

It appears that the significant rise in income per capita in the Post-Malthusian Regime increased the desirable number of surviving offspring and thus, despite the decline in mortality rates, fertility increased significantly so as to enable households to reach this higher desirable level of surviving offspring.

Fertility Rates and Marriage Age

Fertility was controlled during this period, despite the absence of modern contraceptive methods,

[19] See Dyson and Murphy (1985) as well.

partly via adjustment in marriage rates.[20] As depicted in Figure 2.13, increased fertility was achieved by earlier female's age of marriages and a decline in fertility by a delay in the marriage age. The same pattern is observed in the relationship between *Crude Birth Rates* and Female's age of marriages, or alternatively *Crude Marriage Rates* (per 1000).

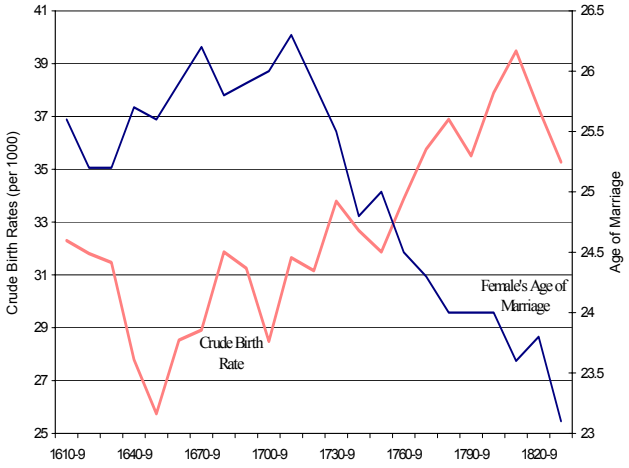


Figure 2.13. Fertility Rates and Female's Age of Marriage
Source: Wrigley and Schofield (1983)

2.2.3 Industrialization and Urbanization

The take-off of developed and less developed regions from the Malthusian epoch was associated with the acceleration in the process of industrialization as well as with a significant rise in urbanization.

Industrialization

The take-off in the developed regions was accompanied by a rapid process of industrialization. As depicted in Figure 2.14, Per-Capita Level of Industrialization (measuring per capita volume of industrial production) increased significantly in the United Kingdom since 1750, rising 50% over the 1750-1800 period, quadrupling in the years 1800-1860, and nearly doubling in the time period 1860-1913. Similarly per-capita level of industrialization accelerated in the United States, doubling in the 1750-1800 as well as 1800-1860 periods , and increasing six-fold in the years 1860-1913. A similar pattern was experienced by Germany, France, Sweden, Switzerland, Belgium, and Canada as of 1800, and industrialization nearly doubled in the 1800-1860 period, further accelerating in the time interval 1860-1913.

The take-off of less developed economies in the 20th century was associated with increased industrialization as well. However, as depicted in Figure 2.14, during the 19th century these economies experienced a decline in *per capita industrialization* (i.e., per capita volume of industrial production), reflecting the adverse effect of the sizable increase in population on the level of industrial production per capita (even in the absence of an absolute decline in industrial production) as well as the forces of globalization and colonialism, that induced less developed economies to specialize in the production of raw materials.[21]

[20] This mechanism is reflected in the assertion of William Cobbett (1763 - 1835) – a leader of the campaign against the changes brought by the Industrial Revolution – *"... men, who are able and willing to work, cannot support their families, and ought... to be compelled to lead a life of celibacy, for fear of having children to be starved."*

[21] The sources of the decline in the industrialization of less developed economies is explored by Galor and Mountford (2003). The effect of colonialism on the patterns of production and thus trade is examined by Acemoglu, Johnson and

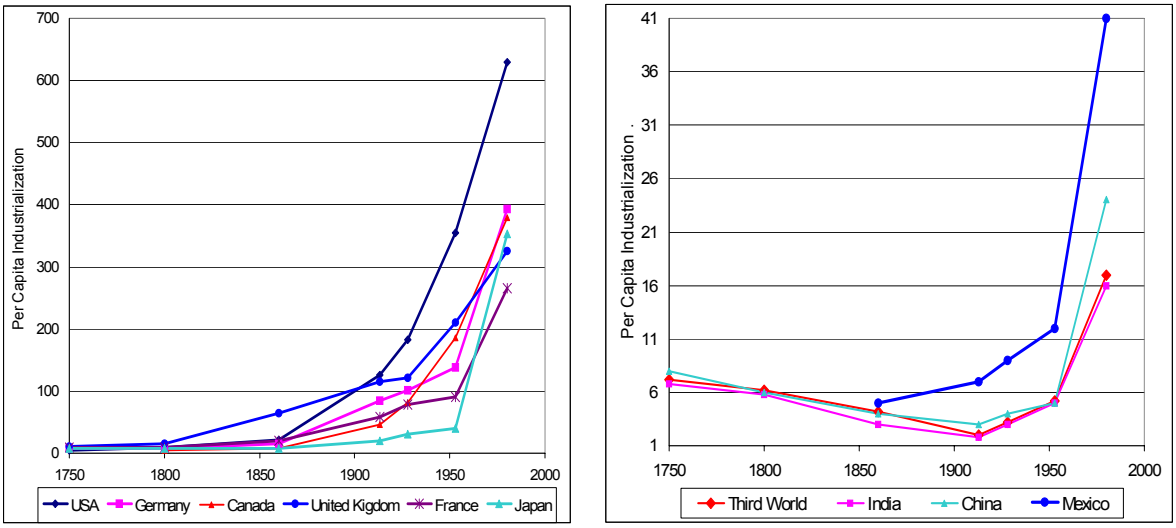


Figure 2.14. Per Capita Levels of Industrialization: (UK in 1900=100)

Source: Bairoch (1982)[22]

Urbanization

The take-off from Malthusian stagnation and the acceleration in the process of industrialization increased significantly the process of urbanization. As depicted in Figure 2.15, the percentage of the population that lived in European cities with a population larger than 10,000 people nearly tripled over the years 1750-1870, from 17% to 54%. Similarly, the percentage of the population in England that lived in cities with population larger than 5,000 quadrupled over the 1750-1910 period, from 18% to 75% (Bairoch 1988)).

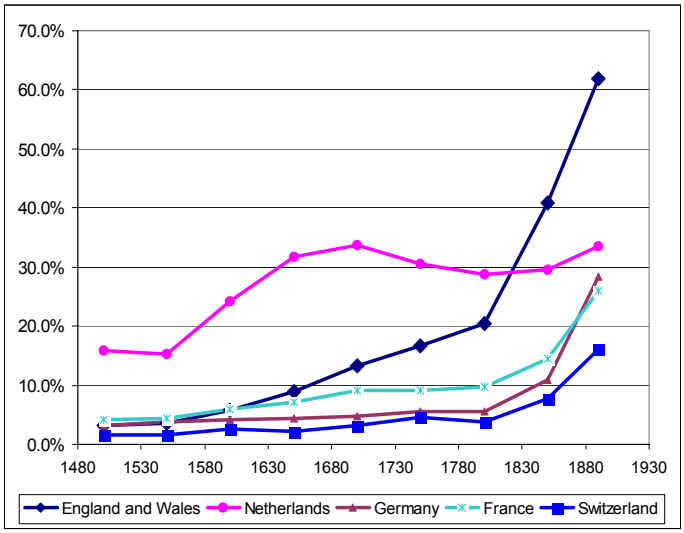


Figure 2.15. Percentage of the Population in Cities with Population larger than 10,000
Sources: Bairoch (1988) and de Vries (1984)

Robinson (2001) and Bertocchi and Canova (2002).

[22]Notes: Countries are defined according to their 1913 boundaries. Germany from 1953 is defined as East and West Germany. India after 1928 includes Pakistan.

This rapid processes of industrialization and urbanization was accompanied by a rapid decline in the share of agricultural production in total output, as depicted in Figure 2.16. For instance, this share declined in England from 40% in 1790 to 7% in 1910.

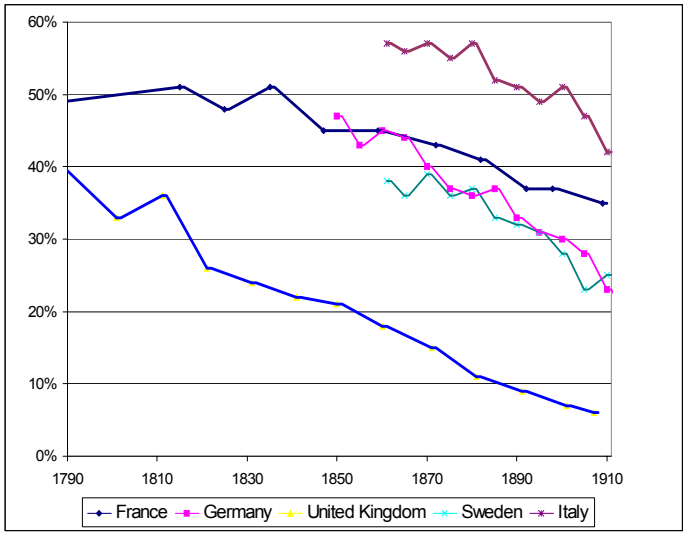


Figure 2.16. The Decline in the Percentage of Agricultural Production in Total Output: Europe: 1790-1910
Source: Mitchell (1981)

2.2.4 Early Stages of Human Capital Formation

The acceleration in technological progress during the Post-Malthusian Regime and the associated increase in income per capita stimulated the accumulation of human capital in the form of literacy rates, schooling, and health. The increase in the investment in human capital was induced by the gradual relaxation in the households budget constraints (as reflected by the rise in real wages and income per capita), as well as by qualitative changes in the economic environment that increased the demand for human capital and induced households to invest in the education of their offspring.

In the first phase of the Industrial Revolution, human capital had a limited role in the production process. Education was motivated by a variety of reasons, such as religion, enlightenment, social control, moral conformity, sociopolitical stability, social and national cohesion, and military efficiency. The extensiveness of public education was therefore not necessarily correlated with industrial development and it differed across countries due to political, cultural, social, historical and institutional factors. In the second phase of the Industrial Revolution, however, the demand for education increased, reflecting the increasing skill requirements in the process of industrialization.[23]

During the post-Malthusian regime, the average number of years of schooling in England and Wales rose from 2.3 for the cohort born between 1801 and 1805 to 5.2 for the cohort born in the year 1852-1856 (Matthews et al., 1982). Furthermore, human capital as reflected by the level of health of the labor force increased over this period. In particular, between 1740 and 1840 life expectancy at birth rose from 33 to 40 in England (Figure 2.8), and from 25 to 40 in France.

The process of industrialization was ultimately characterized by a gradual increase in the relative importance of human capital in less developed economies as well. As documented by Barro and

[23]Evidence suggests that in Western Europe, the economic interests of capitalists were a significant driving force behind the implementation of educational reforms, reflecting the interest of capitalists in human capital formation and thus in the provision of public education [Galor and Moav (2004)].

Lee (2000) educational attainment increased significantly across all less developed regions in the Post-Malthusian Regime (that ended with the decline in population growth in the 1970s in Latin America and Asia, and was still in motion in Africa at the end of the 20th century). In particular, the average years of schooling increased from 3.5 in 1960 to 4.4 in 1975, in Latin America, from 1.6 in 1960 to 3.4 in 2000 in Sub-Saharan Africa, and from 1.4 in 1960 to 1.9 in 1975 in South Asia.

2.3 The Sustained Growth Regime

The acceleration in technological progress and industrialization in the Post-Malthusian Regime and its interaction with the accumulation of human capital brought about a demographic transition, paving the way to a transition to an era of sustained economic growth. In the post demographic-transition period, the rise in aggregate income due to technological progress and factor accumulation has no longer been counterbalanced by population growth, permitting sustained growth in income per capita in regions that have experienced sustained technological progress and factor accumulation.

The transition of the developed regions of Western Europe and the Western Offshoots to the state of sustained economic growth occurred towards the end of the 19th century, whereas the transition of the less developed regions of Asia and Latin America occurred towards the end of the 20th century. Africa, in contrast, is still struggling to make this transition.

2.3.1 Growth of Income Per Capita

During the Sustained Growth Regime the average growth rate of output per capita increased significantly in association with the decline in population growth. As depicted in Figure 2.11, the decline in population growth in Western Europe as well as the Western Offshoots was followed by a significant increase in income per capita and in many of the less advanced economies a significant increase in income per capita was followed by a demographic transition.

Income per capita in the last century has advanced at a stable rate of about 2% per year in Western Europe and the Western Offshoots, as depicted in Figure 2.17.

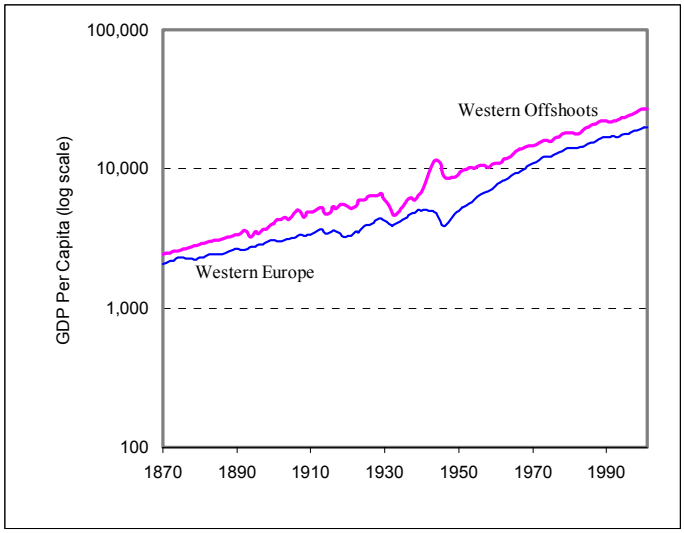


Figure 2.17. Sustained Economic Growth: Western Europe and the Western Offshoots, 1870-2001
Source: Maddison (2003)

In contrast, less developed regions experienced a sustained growth rate of output per capita only in the last decades. As depicted in Figure 2.18, the growth rate of output per capita in Asia has been stable in the last 50 years, the growth rate in Latin America has been declining over this period, and the growth of Africa vanished in the last few decades.[24]

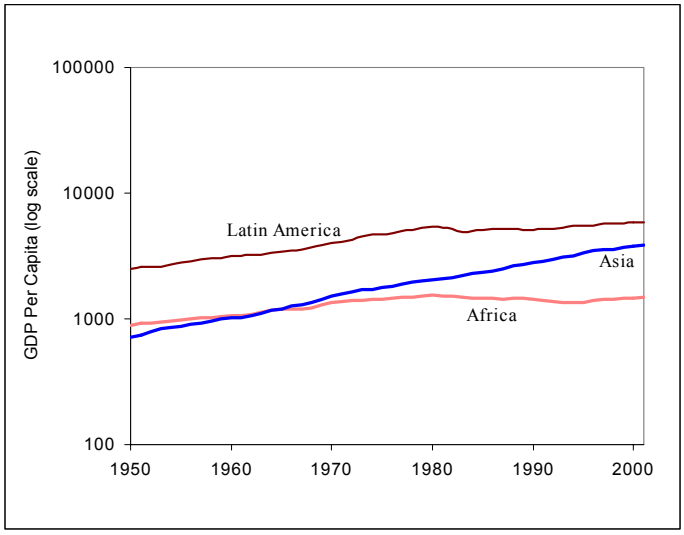


Figure 2.18. Income Per Capita in Africa, Asia and Latin America, 1950-2001
Source: Maddison (2003)

The transition to a state of sustained economic growth in developed as well as less developed regions was accompanied by a rapid process of industrialization. As depicted in Figure 2.14, the *Per Capita Level of Industrialization* (measuring per capita volume of industrial production) doubled in the time period 1860-1913 and tripled in the course of the 20th century. Similarly the per capita level of industrialization in the United States, increased six-fold over the years 1860-1913, and tripled along the 20th century. A similar pattern was experienced by Germany, France, Sweden, Switzerland, Belgium, and Canada where industrialization increased significantly in the time interval 1860-1913 as well as over the rest of the 20th century. Moreover, less developed economies that made the transition to a state of sustained economic growth in recent decades have experienced a significant increase in industrialization.

The transition to a state of sustained economic growth was characterized by a gradual increase in the importance of the accumulation of human capital relative to physical capital as well as with a sharp decline in fertility rates. In the first phase of the Industrial Revolution (1760-1830), capital accumulation as a fraction of GDP increased significantly whereas literacy rates remained largely unchanged. Skills and literacy requirements were minimal, the state devoted virtually no resources to raise the level of literacy of the masses, and workers developed skills primarily through on-the-job training (Green 1990, and Mokyr 1990, 1993). Consequently, literacy rates did not increase during the period 1750-1830 (Sanderson 1995)). As argued by Landes (1969, p. 340) "although certain workers - supervisory and office personal in particular - must be able to read and do the elementary arithmetical operations in order to perform their duties, large share of the work of industry can be performed by illiterates as indeed it was especially in the early days of the Industrial Revolution."

In the second phase of the Industrial Revolution, however, capital accumulation subsided, the education of the labor force markedly increased and skills became necessary for production. The investment ratio which increased from 6% in 1760 to 11.7% in the year 1831, remained at around 11%

[24]Extensive evidence about the growth process in the last four decades is surveyed by Barro and Sala-i-Martin (2003).

on average in the years 1856-1913 (Crafts 1985 and Matthews et al. 1982). In contrast, the average years of schooling of the male labor force which did not change significantly until the 1830s, tripled by the beginning of the 20th century (Matthews et al. 1982, p 573). The significant rise in the level of income per capita in England as of 1865, as depicted in Figure 2.19, was associated with an increase in the standard of living (Voth (2004), and an increase in school enrollment of 10-year olds from 40% in 1870 to 100% in 1900. Moreover, Total fertility Rates in England sharply declined over this period from about 5 in 1875, to nearly 2 in 1925.

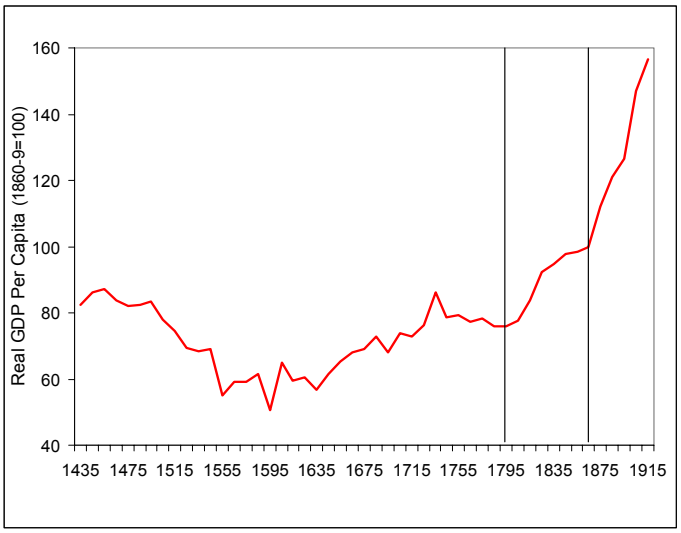


Figure 2.19. The Sharp Rise in Real GDP Per Capita in the transition to Sustained Economic Growth: England 1870-1915
Source: Clark (2001) and Feinstein (1972)

The transition to a state of sustained economic growth in the US, as well, was characterized by a gradual increase in the importance of the accumulation of human capital relative to physical capital. Over the time period 1890-1999 the contribution of human capital accumulation to the growth process in the US nearly doubled whereas the contribution of physical capital declined significantly. Goldin and Katz (2001) show that the rate of growth of educational productivity was 0.29% per year over the 1890-1915 period, accounting for about 11% of the annual growth rate of output per capita over this period.[25] In the period 1915-1999, the rate of growth of educational productivity was 0.53% per year accounting for about 20% of the annual growth rate of output per capita over this period. Abramovitz and David (2000) report that the fraction of the growth rate of output per capita that is directly attributed to physical capital accumulation declined from an average of 56% in the years 1800-1890 to 31% in the period 1890-1927 and 21% in the time interval 1929-1966.

2.3.2 The Demographic Transition

The demographic transitions swept the world in the course of the last century. The unprecedented increase in population growth during the Post-Malthusian regime was ultimately reversed and the demographic transition brought about a significant reduction in fertility rates and population growth in various regions of the world, enabling economies to convert a larger share of the fruits of factor accumulation and technological progress into growth of income per capita. The demographic transition enhanced the growth process via three channels: (a) Reduction of the dilution of the stock of capital

[25] They measure educational productivity by the contribution of education the educational wage differentials.

and land. (b) Enhancement of the investment in the human capital of the population. (c) Alteration of the age distribution of the population which temporarily increased the size of the labor force relative to the population as a whole.

The Decline in Population Growth

The evolution of population growth in the world economy, as depicted in Figure 2.5, has been non-monotonic. The growth of world population was sluggish during the Malthusian epoch, creeping at an average annual rate of about 0.1% over the years 0-1820. The Western European take-off along with that of the Western Offshoots brought about a sharp increase in population growth in these regions and consequently in the world as a whole. The annual average rate of population growth in the world increased gradually reaching 0.8% in the years 1870-1913. The delayed take-off of less developed regions and the significant increase in their income per capita generated a further gradual increase in the rate of population growth in the world, despite the decline in population growth in Western Europe and the Western Offshoots, reaching a high level of 1.92% per year in the period 1950-1973, . Ultimately, however, the onset of the demographic transition in less developed economies in the second half of the 20th century, reduced population growth to an average rate of about 1.63% per year in the 1973-1998 period.

The timing of the demographic transition differed significantly across regions. As depicted in Figure 2.20, the reduction in population growth occurred in Western Europe, the Western Offshoots, and Eastern Europe towards the end of the 19th century and in the beginning of the 20th century, whereas Latin America and Asia experienced a decline in the rate of population growth only in the last decades of the 20th century. Africa's population growth, in contrast, has been rising steadily, although this pattern is likely to reverse in the near future due to the decline in total fertility rate in this region since the 1980s.

The Western Offshoots experienced the earliest decline in population growth, from an average annual rate of 2.87% in the period 1820-1870 to an annual average rate of 2.07% in the time interval 1870-1913 and 1.25% in the years 1913-1950.[26] In Western Europe population growth declined from a significantly lower average level of 0.77% per year in the period 1870-1913 to an average rate of 0.42%

[26]Migration played a significant role in the rate of population growth of these land-abundant countries.

per year in the 1913-1950 period. A similar reduction occurred in Eastern Europe as well.[27]

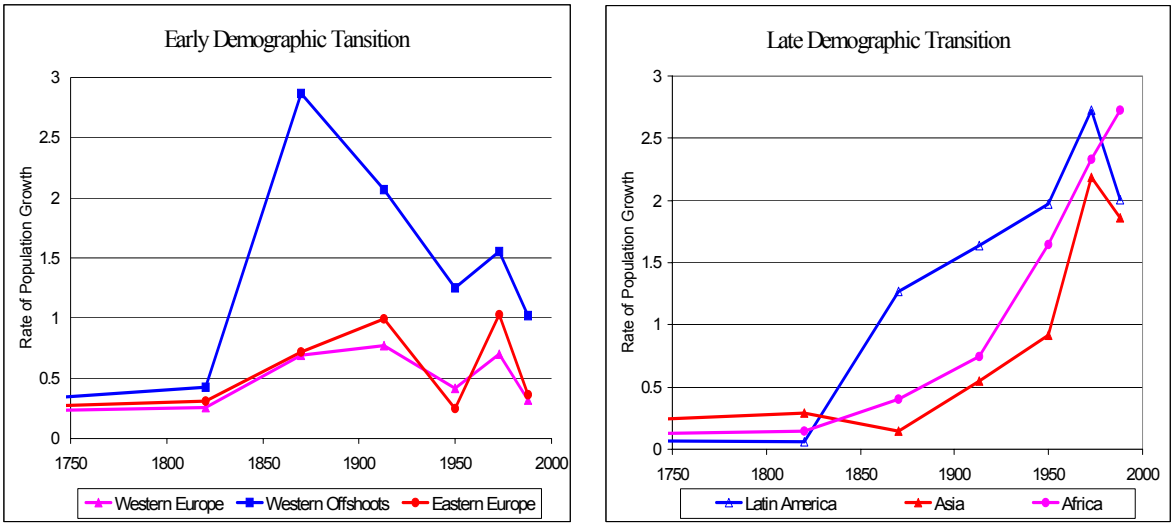


Figure 2.20. The Differential Timing of the Demographic Transition Across Regions
Source: Maddison (2001)

In contrast, in Latin America and Asia the reduction in population growth started to take place in the 1970s, whereas the average population growth in Africa has been rising, despite a modest decline in fertility rates.[28] Latin America experienced a decline in population growth from an average annual rate of 2.73% in the years 1950-1973 to an annual average rate of 2.01% in the period 1973-1998 Similarly, Asia (excluding Japan) experienced a decline in population growth from an average annual rate of 2.21% in the time period 1950-1973 to an average annual rate of 1.86% in the 1973-1998 period. Africa's increased resources in the Post-Malthusian Regime, however, has been channeled primarily towards population growth.

Africa's population growth rate has increased monotonically from a modest average annual rate of 0.4% over the years 1820-1870, to a 0.75% in the time interval 1870-1913, 1.65% in the period 1913-1950, 2.33% in 1950-1973, and a rapid average annual rate of 2.73% in the 1973-1998 period . Consequently, the share of the African population in the world increased by 41% in the 60 year period, 1913-1973 (from 7% in 1913 to 9.9% in 1973), and an additional 30% in the last 25 years, from 9.9% in 1973 to 12.9% in 1998. The decline in fertility in less developed regions, however, has been more significant, indicating a sharp forthcoming decline in population growth in the next decades.

Fertility Decline

The decline in population growth stem from a decline in fertility rates. As depicted in Figure 2.21, *Total Fertility Rate* over the period 1960-1999 plummeted from 6 to 2.7 in Latin America and declined sharply from 6.14 to 3.14 in Asia.[29] Furthermore, *Total Fertility Rate* in Western Europe and the Western Offshoots declined over this period below the replacement level: from 2.8 in 1960 to 1.5 in 1999 in Western Europe and from 3.84 in 1960 to 1.83 in 1999 in the Western Offshoots. (World

[27] A sharper reduction in population growth occurred in the United Kingdom, from 0.87% per year in the period 1870-1913 to 0.27% per year in the period 1913-1950.

[28] As depicted in Figure 2.18, the decline in Total Fertility Rate in these countries started earlier. The delay in the decline in population growth could be attributed to an increase in life expectancy as well as an increase in the relative size of cohorts of women in a reproduction age.

[29] For a comprehensive discussion of the virtues and drawbacks of the various measures of fertility: *TFR*, *NNR*, and *CBR*, see Weil (2004).

Development Indicators, 2001). Even in Africa the *Total Fertility Rate* declined moderately from 6.55 in 1960 to 5.0 in 1999.

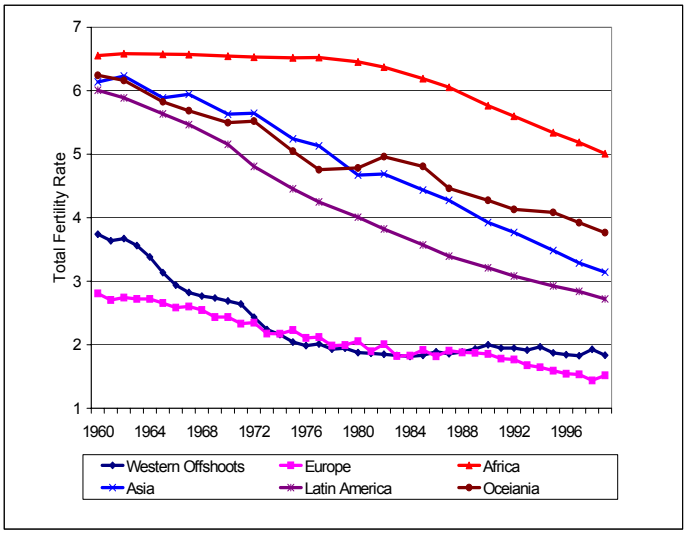


Figure 2.21. The Evolution of Total Fertility Rate Across Regions, 1960-1999
Source: World Development Indicators (2001).

The demographic transition in Western Europe occurred towards the turn of the 19th century. A sharp reduction in fertility took place simultaneously in several countries in the 1870s, and resulted in a decline of about 1/3 in fertility rates in various states within a 50 year period.[30]

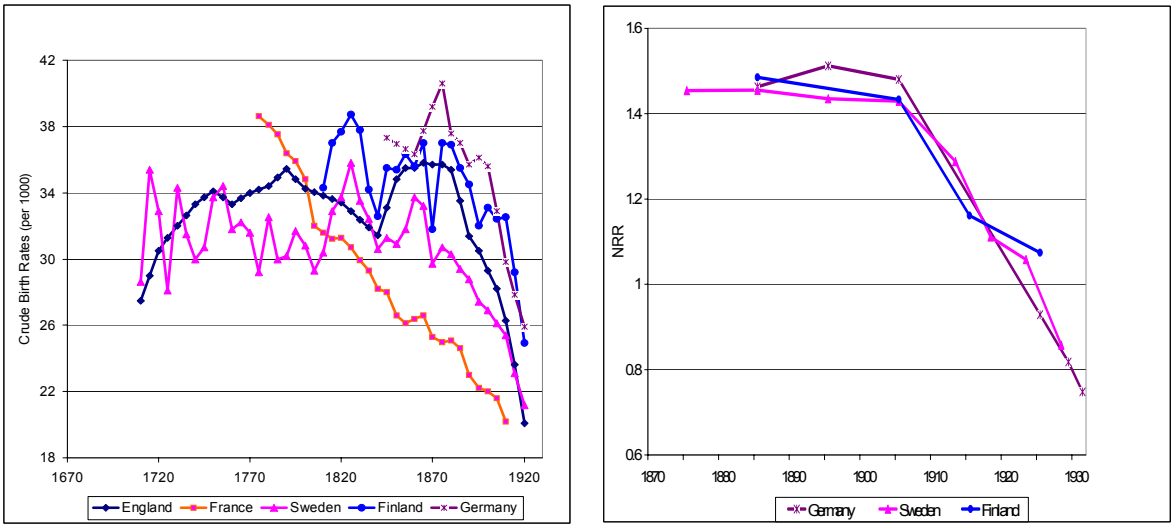


Figure 2.22. The Demographic Transition in Western Europe:
Crude Birth Rates and *Net Reproduction Rates*
Source: Andorka (1978) and Kuzynski (1969)

As depicted in Figure 2.22, *Crude Birth Rates* in England declined by 44%, from 36 (per 1000) in 1875, to 20 (per 1000) in 1920. Similarly, live births per 1000 women aged 15-44 fell from 153.6 in

[30]Coale and Treadway (1986) find that a 10% decline in fertility rates was completed in 59% of all European countries in the time period 1890-1920. In particular, a 10% decline was completed in Belgium in 1881, Switzerland in 1887, Germany in 1888, England and Wales in 1892, Scotland in 1894, Netherlands in 1897, Denmark in 1898, Sweden in 1902, Norway in 1903, Austria in 1907, Hungary in 1910, Finland in 1912, Greece and Italy in 1913, Portugal in 1916, Spain 1920, and Ireland in 1922.

1871-80 to 109.0 in 1901-10 (Wrigley, 1969). In Germany, *Crude Birth Rates* declined 37%, from 41 (per 1000) in 1875 to 26 (per 1000) in 1920. Sweden's *Crude Birth Rates* declined 32%, from 31 (per 1000) in 1875 to 21 (per 1000) in 1920, and in Finland they declined 32%, from 37 in 1875 to 25 (per 1000) in 1920. Finally, although the timing of demographic transition in France represents an anomaly, starting in the second half of the 18th century, France experienced an additional significant reduction in fertility in the time period 1865-1910, where *Crude Birth Rates* declined by 26%, from 27 (per 1000) in 1965 to 20 (per 1000) in 1910.

The decline in the crude birth rates in the course of the demographic transition was accompanied by a significant decline in the *Net Reproduction Rate (*i.e., the number of daughters per woman who reach the reproduction age), as depicted in Figure 2.22. Namely, the decline in fertility during the demographic transition outpaced the decline in mortality rates, and brought about a decline in the number of children who survived to their reproduction age.

Similar patterns are observed in the evolution of *Total Fertility Rates* in Western Europe, as depicted in Figure 2.23. *Total Fertility Rates* (TFR) peaked in the 1870s and then decline sharply and simultaneously across Western European States. In England, TFR declined by 51%, from 4.94 children in 1875, to 2.4 in 1920. In Germany, TFR declined 57%, from 5.29 in 1885 to 2.26 in 1920. Sweden's TFR declined 61%, from 4.51 in 1876 to 1.77 in 1931, in Finland they declined 52%, from 4.96 in 1876 to 2.4 in 1931 and in France where a major decline occurred in the years 1750-1850, additional decline took place in the same time period from 3.45 in 1980 to 1.65 in 1920.

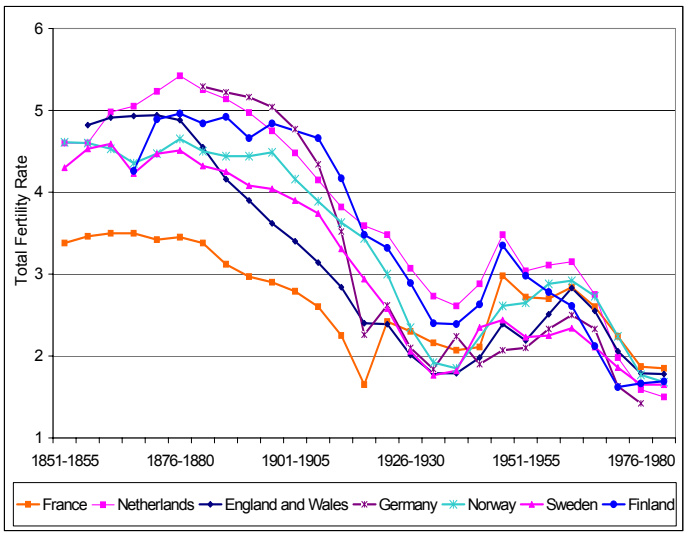


Figure 2.23. The Demographic Transition in Western Europe: Total Fertility Rates
Source: Chesnais (1992)

Mortality Decline

The mortality decline which preceded the decline in fertility rates in most countries in the world, with the notable exceptions of France and the United States, has been, unjustifiably, viewed by demographers as the prime force behind the demographic transition. The evidence provided in section 3.3.1, suggests that this viewpoint is inconsistent with historical evidence.

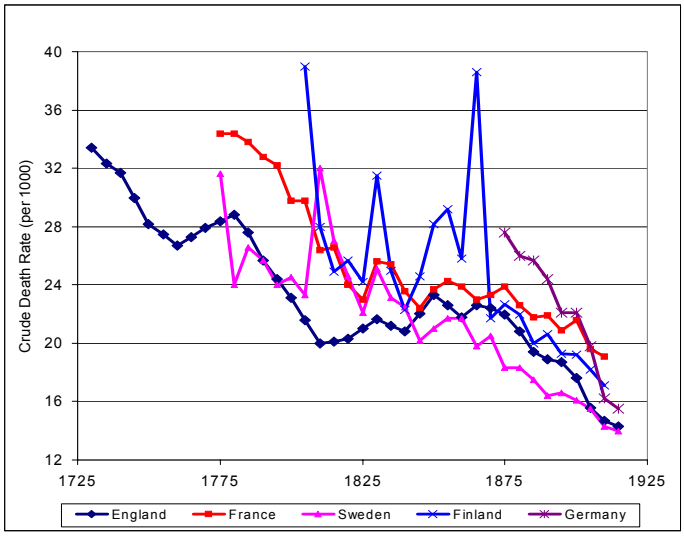


Figure 2.24. The Mortality Decline in Western Europe, 1730-1920
Source: Andorka (1978)

The decline in mortality rates preceded the decline in fertility rates in Western European countries in the 1730-1920 period, as depicted in Figures 2.22 and 2.24. The decline in mortality rates began in England 140 years prior to the decline in fertility and in Sweden and Finland nearly 100 years prior to the decline in fertility.

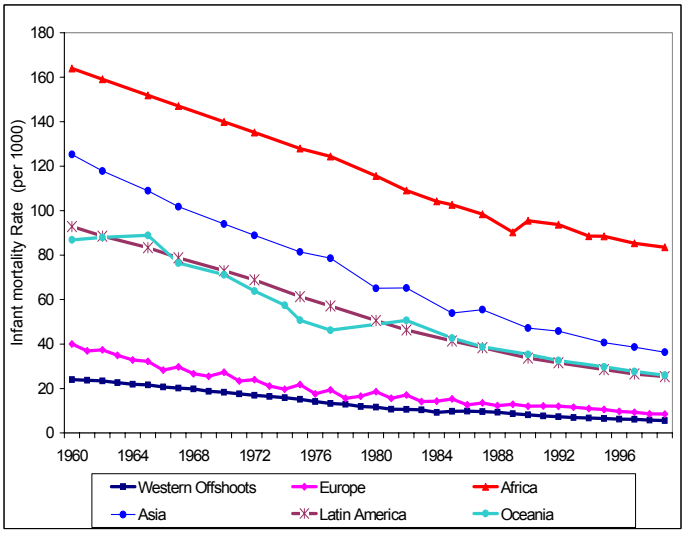


Figure 2.25. The Decline in Infant Mortality Rates Across Regions, 1960-1999
Source: World Development Indicators (2001).

A similar sequence of events emerges from the pattern of mortality and fertility decline in less developed regions. As depicted in Figures 2.21 and 2.25, a sharp decline in infant mortality rates as of 1960 preceded the decline in fertility rates in Africa that took place in 1980. Moreover, the existing evidence shows a simultaneous reduction in mortality and fertility in the 1960-2000 period in all other regions.[31]

[31]Extrapolation about mortality rates prior to 1960 suggests that a similar pattern appears in Asia and Latin America.

Life Expectancy

The decline in mortality rates in developed countries since the 18th century, as depicted in Figure 2.24, corresponded to a gradual increase in life expectancy generating a further inducement for investment in human capital. As depicted in Figure 2.26, life expectancy at birth in England increased at a stable pace from 32 years in the 1720s to about 41 years in the 1870s. This pace of the rise in life expectancy increased towards the end of the 19th century and life expectancy reached the levels of 50 years in the year 1906, 60 years in the year 1930 and 77 years in the year 1996.

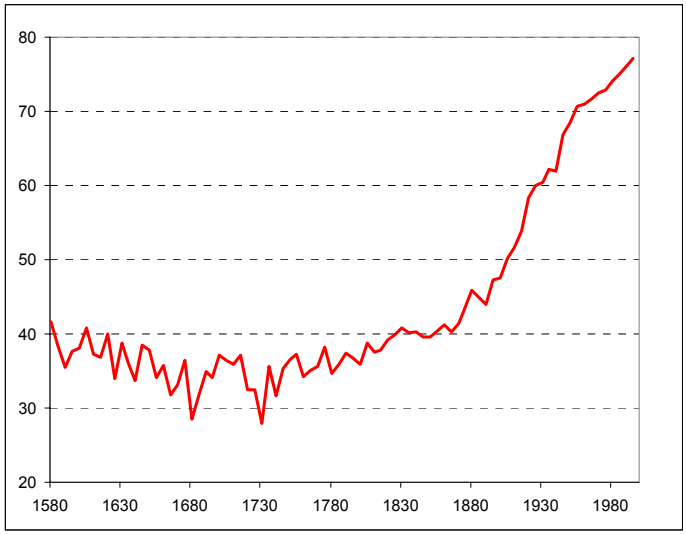


Figure 2.26. The Evolution of Life Expectancy: England 1726-1996
Source: Wrigley and Schofield (1981) for 1726-1871 and Human Mortality Database (2003) for 1876-1996

Similarly, the significant decline in mortality rates across the developed regions in the last two centuries and across less developed regions in the past century, corresponded to an increase in life expectancy. As depicted in Figure 2.27, life expectancy increased significantly in developed regions in the 19th century, whereas the rise in life expectancy in less developed regions occurred throughout the 20th century, stimulating further human capital formation.

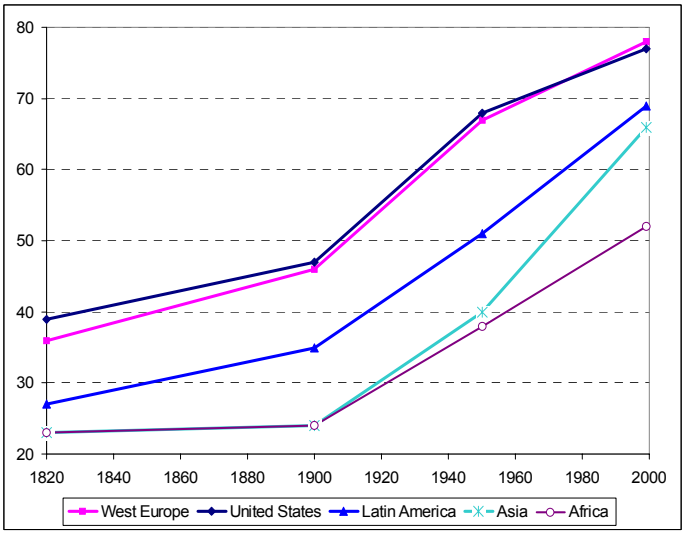


Figure 2.27. The Evolution of Life Expectancy Across Regions, 1820-1999
Source: Maddison (2001).

In particular, life expectancy nearly tripled in the course of the 20th century in Asia, rising from a level of 24 years in 1960 to 66 years in 1999, reflecting the rise in income per capita as well as the diffusion of medical technology. Similarly, life expectancy in Africa doubled from 24 years in 1900 to 52 years in 1999. The more rapid advancement in income per capita in Latin America generated an earlier rise in life expectancy, and life expectancy increased modestly during the 19th century and more significantly in the course of the 20th century, from 35 years in 1900 to 69 years in 1999.

2.3.3 Industrial Development and Human Capital Formation

The process of industrialization was characterized by a gradual increase in the relative importance of human capital for the production process. The acceleration in the rate of technological progress increased gradually the demand for human capital, inducing individuals to invest in education, and stimulating further technological advancement. Moreover, in developed as well as less developed regions the onset of the process of human capital accumulation preceded the onset of the demographic transition, suggesting that the rise in the demand for human capital in the process of industrialization and the subsequent accumulation of human capital played a significant role in the demographic transition and the transition to a state of sustained economic growth.

Developed Economies[32]

As observed by Abramowitz (1993 p. 224), "In the nineteenth century, technological progress was heavily biased in a physical capital-using direction...the bias shifted in an intangible (human and knowledge) capital-using direction and produced the substantial contribution of education and other intangible capital accumulation to this century productivity growth." Furthermore, as argued by Goldin (2001), "The modern concept of the wealth of nations emerged by the early twentieth century. It was that capital embodied in the people — human capital — mattered."

In the first phase of the Industrial Revolution, the extent of public education was not correlated with industrial development and it differed across countries due to political, cultural, social, historical and institutional factors. Human capital had a limited role in the production process and education served religious, social, and national goals. In contrast, in the second phase of the Industrial Revolution the demand for skilled labor in the growing industrial sector markedly increased, human capital formation was designed primarily to satisfy the increasing skill requirements in the process of industrialization, and industrialists became involved in shaping the education system.

Notably, the reversal of the Malthusian relation between income and population growth during the demographic transition, corresponded to an increase in the level of resources invested in each child. For example, the literacy rate among men, which was stable at around 65% in the first phase of the Industrial Revolution, increased significantly during the second phase, reaching nearly 100% at the end of the 19th century (Clark 2003), and the proportion of children aged 5 to 14 in primary schools increased significantly in the second half of the 19th century, from 11% in 1855 to 74% in 1900. A similar pattern is observed in other European societies (Flora et al. 1983). In particular, as depicted in Figure 2.28, the proportion of children aged 5 to 14 in primary schools in France increased significantly in the second half of the 19th century, from 30% in 1832 to 86% in 1901.

[32]This section is closely based on the research of Galor and Moav (2004b).

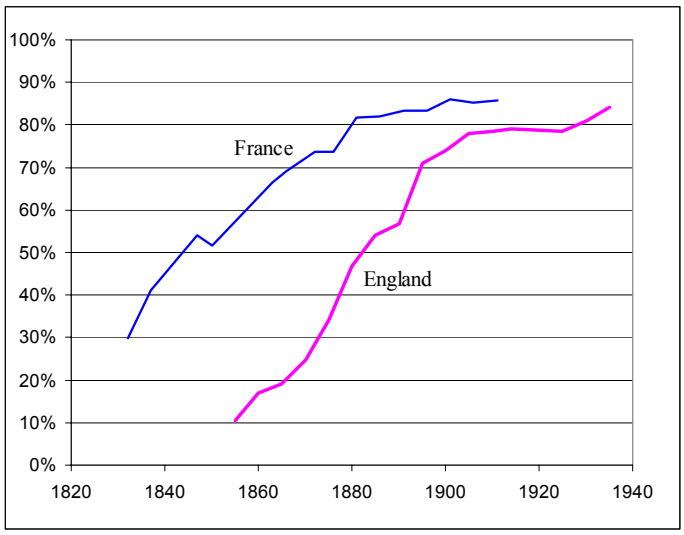


Figure 2.28. The Fraction of Children Age 5-14 in Public Primary Schools, 1820-1940
Source: Flora et al. (1983)

Evidence about the evolution of the return to human capital over this period are scarce and controversial. They do not indicate that the skill premium increased markedly in Europe over the course of the 19th century (Clark 2003). One can argue that the lack of clear evidence about the increase in the return to human capital over this period is an indication for the absence of a significant increase in the demand for human capital. This partial equilibrium argument, however is flawed. The return to human capital is an equilibrium price that is affected both by the demand and the supply of human capital. Technological progress in the second phase of the Industrial Revolution brought about an increase in the demand for human capital, and indeed, in the absence of a supply response, one would have expected an increase in the return to human capital. However, the significant increase in schooling that took place in the 19th century, and in particular the introduction of public education that lowered the cost of education, generated a significant increase in the supply of educated workers. Some of this supply response was a direct reaction of the increase in the demand for human capital, and thus may only operate to partially offset the increase in the return to human capital. However, the removal of the adverse effect of credit constraints on the acquisition of human capital (as reflected by the introduction of public education) generated an additional force that increased the supply of educated labor and operated towards a reduction in the return to human capital.[33]

A. The Industrial Base for Education Reforms in the 19th Century

Education reforms in developed countries in the 18th and the 19th century provide a profound insight about the significance of industrial development in the formation of human capital (and thus in the onset of the demographic transition) in the second half of the 19th century. In particular, the variation in the timing of the establishment of a national system of public education between England and Continental Europe is instrumental in isolating the role of industrial forces in human capital formation from other forces such as social control, moral conformity, enlightenment, sociopolitical stability, social and national cohesion, and military efficiency.

[33]This argument is supported indirectly by contemporary evidence about a higher rate of returns to human capital in less developed economies than in developed economies (Psacharopoulos and Patrinos 2002). The greater prevalence of credit markets imperfections and other barriers for the acquisition of skills in less developed economies generated only a partial supply response to industrial demand for human capital, contributing to this differential in the skill premium.

England In the first phase of the Industrial Revolution (1760-1830), capital accumulation increased significantly without a corresponding increase in the supply of skilled labor. The investment ratio increased from 6% in 1760 to 11.7% in the year 1831 (Crafts 1985, p. 73). In contrast, literacy rates remained largely unchanged and the state devoted virtually no resources to raising the level of literacy of the masses. During the first stages of the Industrial Revolution, literacy was largely a cultural skill or a hierarchical symbol and had limited demand in the production process.[34] For instance, in 1841 only 4.9% of male workers and only 2.2% of female workers were in occupations in which literacy was strictly required (Mitch 1992, pp. 14-15). During this period, an illiterate labor force could operate the existing technology, and economic growth was not impeded by educational retardation.[35] Workers developed skills primarily through on-the-job training, and child labor was highly valuable.

The development of a national public system of education in England lagged behind the continental countries by nearly half a century and the literacy rate hardly increased in the period 1750-1830 (Sanderson 1995, pp. 2-10).[36] As argued by Green (1990, pp. 293-294), "Britain's early industrialization had occurred without direct state intervention and developed successfully, at least in its early stages, within a laissez-faire framework. Firstly, state intervention was thought unnecessary for developing technical skills, where the initial requirements were slight and adequately met by traditional means. Secondly, the very success of Britain's early industrial expansion encouraged a complacency about the importance of scientific skills and theoretical knowledge which became a liability in a later period when empirical knowledge, inventiveness and thumb methods were no longer adequate." Furthermore, as argued by Landes (1969, p. 340) "although certain workers - supervisory and office personnel in particular - must be able to read and do the elementary arithmetical operations in order to perform their duties, large share of the work of industry can be performed by illiterates as indeed it was especially in the early days of the industrial revolution."

England initiated a sequence of reforms in its education system since the 1830s and literacy rates gradually increased. The process was initially motivated by a non-industrial reasons such as religion, social control, moral conformity, enlightenment, and military efficiency, as was the case in other European countries (e.g., Germany, France, Holland, Switzerland) that had supported public education much earlier.[37] However, in light of the modest demand for skills and literacy by the capitalists, the level of governmental support was rather small.[38]

In the second phase of the Industrial Revolution, the demand for skilled labor in the growing industrial sector markedly increased and the proportion of children aged 5 to 14 in primary schools increased from 11% in 1855 to 25% in 1870 (Flora et al. 1983). Job advertisements, for instance, suggest that literacy became an increasingly desired characteristic for employment as of the 1850s (Mitch 1993, p. 292). In light of the industrial competition from other countries, capitalists started to recognize the importance of technical education for the provision of skilled workers. As noted by Sanderson 1995, pp. 10-13), "reading ...enabled the efficient functioning of an urban industrial society laced with letter writing, drawing up wills, apprenticeship indentures, passing bills of exchange, and notice and

[34]See Mokyr (1993, 2001).

[35]Some have argued that the low skill requirements even declined over this period. For instance, Sanderson (1995, p. 89) suggests that "One thus finds the interesting situation of an emerging economy creating a whole range of new occupations which require even less literacy and education than the old ones."

[36]For instance, in his parliamentary speech in defense of his 1837 education bill, the Whig politician, Henry Brougham, reflected upon this gap: "It cannot be doubted that some legislative effort must at length be made to remove from this country the opprobrium of having done less for education of the people than any of the more civilized nations on earth" (Green (1990, pp.10-11)).

[37]The proximity of the education acts in the UK to major wars suggests that the provision of public education was partly a compensation for the services of soldiers.

[38]Even in 1869 the government funded only one-third of school expenditure (Green, 1990, pp. 6-7).

advertisement reading." Moreover, manufacturers argued that: "universal education is required in order to select, from the mass of the workers, those who respond well to schooling and would make a good foreman on the shop floor" (Simon 1987, p. 104).

As it became apparent that skills were necessary for the creation of an industrial society, replacing previous ideas that the acquisition of literacy would make the working classes receptive to radical and subversive ideas, capitalists lobbied for the provision of public education for the masses.[39] The pure laissez-faire policy failed in developing a proper educational system and capitalists demanded government intervention in the provision of education. As James Kitson, a Leeds iron-master and an advocate of technical education explained to the Select Committee on Scientific Instruction (1867-1868): "...the question is so extensive that individual manufacturers are not able to grapple with it, and if they went to immense trouble to establish schools they would be doing it in order that others may reap the benefit" (Green, 1990, p. 295).[40]

An additional turning point in the attitude of capitalists towards public education was the Paris Exhibition of 1867, where the limitations of English scientific and technical education became clearly evident. Unlike the 1851 exhibition in which England won most of the prizes, the English performance in Paris was rather poor; of the 90 classes of manufacturers, Britain dominated only in 10. Lyon Playfair, who was one of the jurors, reported that: "a singular accordance of opinion prevailed that our country has shown little inventiveness and made little progress in the peaceful arts of industry since 1862." This lack of progress "upon which there was most unanimity conviction is that France, Prussia, Austria, Belgium and Switzerland possess good systems of industrial education and that England possesses none" (Green 1990, p. 296).[41]

In 1868, the government established the Parliamentary Select Committee on Scientific Education. This was the origin of nearly 20 years of various parliamentary investigations into the relationship between science, industry, and education, that were designed to address the capitalists' outcry about the necessity of universal public education. A sequence of reports by the committee in 1868, the Royal Commission on Scientific Instruction and the Advancement of Science during the period 1872-75, and by the Royal Commission on Technical Education in 1882, underlined the inadequate training for supervisors, managers and proprietors, as well as workers. They argued that most managers and proprietors did not understand the manufacturing process and thus, failed to promote efficiency, investigate innovative techniques or value the skills of their workers (Green 1990, pp. 297-298). In particular, W. E. Forster, the Vice President of the committee of the Council of Education told The House of Commons: "Upon the speedy provision of elementary education depends our industrial prosperity...if we leave our work-folk any longer unskilled...they will become overmatched in the competition of the world" (Hurt 1971, pp. 223-224). The reports made various recommendations which highlighted the need to redefine elementary schools, to revise the curriculum throughout the entire school system, particularly with respect to industry and manufacture, and to improve teacher training.

In addition, in 1868, secondary schools were investigated by the Schools Inquiry Commission, which found a very unsatisfactory level for the vast majority of schools that employed untrained teachers and used antiquated methods. Their main proposal was to organize a state inspection of secondary

[39]There was a growing consensus among workers and capitalists about the virtues of reform. The labor union movement was increasingly calling for a national system of non-sectarian education. The National Education League (founded in 1869 by radical Liberals and Dissenters) demanded a free, compulsory, non-sectarian national system of education (Green, 1990, p. 302).

[40]Indeed, the Factory Act of 1802 required owners of textile mills to provide elementary instruction for their apprentices, but the law was poorly enforced (Cameron (1989, p. 216-217)).

[41]Moreover, the Nussey brothers, who had written a report on woolen textiles at the Exhibition, returned to Leeds to start a movement for a Yorkshire College of Science.

schools and to provide efficient education geared towards the specific needs of its consumers. In particular, the Royal Commission on Technical Education of 1882 confirmed that England was being overtaken by the industrial superiority of Prussia, France and the United States and recommended the introduction of technical and scientific education into secondary schools.

It appears that the government gradually yielded to the pressure by capitalists as well as labor unions, as reflected by its increased contributions to elementary as well as higher education. In the 1870 Education Act, the government assumed responsibility for ensuring universal elementary education, although it did not provide either free or compulsory education at the elementary level. The Act created a national provision without an integrated system, where voluntary schools existed beside state schools. In 1880, prior to the significant extension of the franchise of 1884 that made the working class the majority in most industrial countries, education was made compulsory throughout England. The 1889 Technical Instruction Act allowed the new local councils to set up technical instruction committees, and the 1890 Local Taxation Act provided public funds that could be spent on technical education (Green, 1990, p. 299).

School enrollment of 10-year olds increased from 40% in 1870 to 100% in 1900, the literacy rate among men, which was stable at around 65% in the first phase of the Industrial Revolution, increased significantly during the second phase reaching nearly 100% at the end of the 19th century (Clark 2002), and the proportion of children aged 5 to 14 in primary schools increased significantly in the second half of the 19th century, from 11% in 1855 to 74% in 1900 (Flora et al. 1983). Finally, the 1902 Balfour Act marked the consolidation of a national education system and created state secondary schools (Ringer 1979 and Green 1990, p. 6) and science and engineering and their application to technology gained prominence (Mokyr 1990, 2002).

Continental Europe The early development of public education occurred in the western countries of continental Europe (e.g., Prussia, France, Sweden, and the Netherlands) well before the Industrial Revolution. The provision of public education at this early stage was motivated by several goals such as social and national cohesion, military efficiency, enlightenment, moral conformity, sociopolitical stability as well as religious reasons. However, as was the case in England, massive educational reforms occurred in the second half of the 19th century due to the rising demand for skills in the process of industrialization. As noted by Green (1990, pp. 293-294) "In continental Europe industrialization occurred under the tutelage of the state and began its accelerated development later when techniques were already becoming more scientific; technical and scientific education had been vigorously promoted from the center as an essential adjunct of economic growth and one that was recognized to be indispensable for countries which wished to close Britain's industrial lead."

In France, indeed, the initial development of the education system occurred well before the Industrial Revolution, but the process was intensified and transformed to satisfy industrial needs in the second phase of the Industrial Revolution. The early development of elementary and secondary education in the 17th and 18th centuries was dominated by the Church and religious orders. Some state intervention in technical and vocational training was designed to reinforce development in commerce, manufacturing and military efficiency. After the French Revolution, the state established universal primary schools. Nevertheless, enrolment rates remained rather low. The state concentrated on the development of secondary and higher education with the objective of producing an effective elite to operate the military and governmental apparatus. Secondary education remained highly selective, offering general and technical instruction largely to the middle class (Green 1990, pp. 135-137 and 141-142)). Legislative proposals during the National Convention quoted by Cubberley (1920, pp. 514-517) are revealing about the

underlying motives for education in this period: "...Children of all classes were to receive education, physical, moral and intellectual, best adapted to develop in them republican manners, patriotism, and the love of labor...They are to be taken into the fields and workshops where they may see agricultural and mechanical operations going on..."

The process of industrialization in France and the associated increase in the demand for skilled labor, as well as the breakdown of the traditional apprenticeship system, significantly affected the attitude towards education. State grants for primary schools were gradually increased in the 1830s and legislation made an attempt to provide primary education in all regions, extend the higher education, and provide teacher training and school inspections. The number of communities without schools fell by 50% from 1837 to 1850 and as the influence of industrialists on the structure of education intensified, education became more stratified according to occupational patterns (Anderson 1975 p. 15, 31). According to Green 1990, p.157): "[This] legislation...reflected the economic development of the period and thus the increasing need for skilled labor." The eagerness of capitalists for rapid education reforms was reflected by the organization of industrial societies that financed schools specializing in chemistry, design, mechanical weaving, spinning, and commerce (Anderson 1975, p 86, 204).

As was the case in England, industrial competition led industrialists to lobby for the provision of public education. The Great Exhibition of 1851 and the London Exhibition of 1862 created the impression that the technological gap between France and other European nations was narrowing and that French manufacturers ought to invest in the education of their labor force to maintain their technological superiority. Subsequently, the reports on industrial education by commissions established in the years 1862 to 1865 reflected the plea of industrialists for the provision of industrial education on a large scale and for the implementation of scientific knowledge in the industry. "The goal of modern education...can no longer be to form men of letters, idle admirers of the past, but men of science, builders of the present, initiators of the future."[42] (Anderson 1975, p. 194).

Education reforms in France were extensive in the second phase of the Industrial Revolution, and by 1881 a universal, free, compulsory and secular primary school system had been established and technical and scientific education further emphasized. Illiteracy rates among conscripts tested at the age of 20 declined gradually from 38% in 1851-55 to 17% in 1876-80 (Anderson 1975, p. 158)), and the proportion of children aged 5 to 14 in primary schools increased from 51.5% in 1850 to 86% in 1901 (Flora et al. 1983)). Hence, the process of industrialization, and the increase in the demand for skilled labor in the production process, led industrialists to support the provision of universal education, contributing to the extensiveness of education as well as to its focus on industrial needs.

In Prussia, as well, the initial steps towards compulsory education took place at the beginning of the 18th century well before the Industrial Revolution. Education was viewed at this stage primarily as a method to unify the state. In the second part of the 18th century, education was made compulsory for all children aged 5 to 13. Nevertheless, these regulations were not strictly enforced due to the lack of funding associated with the difficulty of taxing landlords for this purpose, and due to the loss of income from child labor. At the beginning of the 19th century, motivated by the need for national cohesion, military efficiency, and trained bureaucrats, the education system was further reformed, establishing provincial and district school boards, making education a secular activity and compulsory for a three-year period, and reconstituting the Gymnasium as a state institution that provided nine years of education for the elite (Cubberly 1920 and Green 1990).

The process of industrialization in Prussia and the associated increase in the demand for skilled labor led to significant pressure for educational reforms and thereby to the implementation of universal

[42] L'Enseignement professionnel, ii (1864), p. 332, quoted in Anderson (1975).

elementary schooling. Taxes were imposed to finance the school system and teacher training and certification were established. Secondary schools started to serve industrial needs as well, and the Realschulen, which emphasized the teaching of mathematics and science, was gradually adopted, and vocational and trade schools were founded. Total enrolment in secondary school increased sixfold from 1870 to 1911 (Flora et al. 1983). "School courses...had the function of converting the occupational requirements of public administration, commerce and industry into educational qualifications..." (Muller 1987, pp. 23-24). Furthermore, the Industrial Revolution significantly affected the nature of education in German universities. German industrialists who perceived advanced technology as the competitive edge that could boost German industry, lobbied for reforms in the operation of universities, and offered to pay to reshape their activities so as to favor their interest in technological training and industrial applications of basic research (McClelland 1980, p. 300-301).

The structure of education in the Netherlands also reflected the interest of capitalists in the skill formation of the masses. In particular, as early as the 1830s, industrial schools were established and funded by private organizations, representing industrialists and entrepreneurs. Ultimately, in the latter part of the 19th century, the state, urged by industrialists and entrepreneurs, started to support these schools (Wolthuis, 1999, pp. 92-93, 119, 139-140, 168, 171-172).

United States The process of industrialization in the US also increased the importance of human capital in the production process. Evidence provided by Abramowitz and David (2000) and Goldin and Katz (2001) suggests that over the period 1890-1999, the contribution of human capital accumulation to the growth process of the United States nearly doubled.[43] As argued by Goldin (2001), the rise of the industrial, business and commerce sectors in the late 19th and early 20th centuries increased the demand for managers, clerical workers, and educated sales personnel who were trained in accounting, typing, shorthand, algebra, and commerce. Furthermore, in the late 1910s, technologically advanced industries demanded blue-collar craft workers who were trained in geometry, algebra, chemistry, mechanical drawing, etc. The structure of education was transformed in response to industrial development and the increasing importance of human capital in the production process, and American high schools adapted to the needs of the modern workplace of the early 20th century. Total enrolment in public secondary schools increased 70-fold from 1870 to 1950. (Kurian, 1994).[44]

B. Human Capital, Factor Prices and Inequality

In the first phase of the Industrial Revolution, prior to the implementation of significant education reforms, physical capital accumulation was the prime engine of economic growth. In the absence of significant human capital formation, the concentration of capital among the capitalist class widened

[43]It should be noted that literacy rates in the US were rather high prior to this increase in the demand for skilled labor. Literacy rates among the white population were already 89% in 1870, 92% in 1890, and 95% in 1910 (Engerman and Sokoloff (2000)). Education in earlier periods was motivated by social control, moral conformity, and social and national cohesion, as well as required skills for trade and commerce. In particular, Field (1976) and Bowles and Gintis (1975) argue that educational reforms are designed to *sustain* the existing social order, by displacing social problems into the school system.

[44]As noted by Galor and Moav (2004), due to differences in the structure of education finance in the US in comparison to European countries, capitalists in the US had only limited incentives to lobby for the provision of education and support it financially. Unlike the central role that government funding played in the provision of public education in European countries, the evolution of the education system in the US was based on local initiatives and funding. The local nature of the education initiatives in the US induced community members, in urban as well as rural areas, to play a significant role in advancing their schooling system. American capitalists, however, faced limited incentives to support the provision of education within a county in an environment where labor was mobile across counties and the benefits from educational expenditure in one county may be reaped by employers in other counties. "The impetus to expand education to the secondary level was primarily a grassroots movement led by parents, employers, and even young people themselves" (Goldin (1999)).

wealth inequality. Once education reforms were implemented, however, the significant increase in the return to labor relative to capital, as well as the significant increase in the real return to labor and the associated accumulation of assets by the workers, brought about a decline in inequality.

Evidence suggests that in the first phase of the Industrial Revolution, prior to the implementation of education reforms, capital accumulation brought about a gradual increase in wages along with an increase in the wage-rental ratio. Education reforms in the second phase of the Industrial Revolution were associated with a sharp increase in real wages along with a sharp increase in the wage-rental ratio.[45] Finally, wealth inequality widened in the first phase of the Industrial Revolution and reversed its course in the second phase, once significant education reforms were implemented.

As documented in Figure 2.29, over the time period 1823-1915, wealth inequality in the UK reached a peak around 1870 and declined thereafter, in close association with the patterns of enrolment rates and factor prices, depicted in Figures 2.28 and 2.29.[46] It appears that the decline in inequality is associated with the significant changes that occurred around 1870 in the relative returns to the main factors of production possessed by capitalists and workers. These changes in factor prices reflect the increase in enrolment rates – in particular the process of education reforms from 1830 to 1870 and its consolidation in the Education Act of 1870 – and its delayed effect on the skill level per worker.

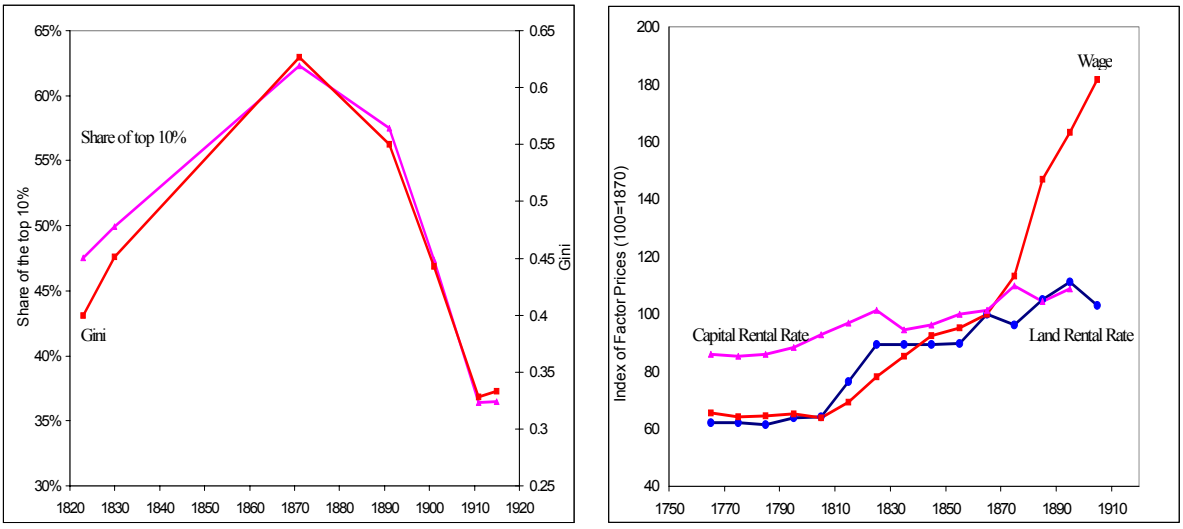


Figure 2.29. Wealth Inequality and Factor Prices: England 1820-1920
Source: Williamson (1985) for inequality and Clark (2002, 2003) for factor prices

Similar patterns of the effect of education on factor prices and therefore on inequality are observed in France as well. As argued by Morrisson and Snyder (2000), wealth inequality in France increased during the first half of the 19th century, and started to decline in the last decades of the 19th century in close association with the rise in education rates depicted in Figure 2.28, the rise in real wages depicted in Figure 2.10, and a declining trend of the return to capital over the 19th century. The decline in inequality in France appears to be associated with the significant changes in the relative returns to the main factors of production possessed by capitalists and workers in the second part of the 19th century As depicted in Figures 2.10, based on the data presented in Levy-Leboyer and Bourguignon (1990), real wages as well as the wage-rental ratio increased significantly as of 1860, reflecting the rise in the demand

[45]It should be noted that the main source of the increase in real wages was not a decline in prices. Over this period nominal wages increased significantly as well.

[46]It should be noted that the return to capital increased moderately over this period, despite the increase in the supply of capital, reflecting technological progress, population growth, and accumulation of human capital.

for skilled labor and the effect of the increase in enrolment rates on the skill level per worker.

The German experience is consistent with this pattern as well. Inequality in Germany peaked towards the end of the 19th century (Morrisson and Snyder 2000) in association with a significant increase in the real wages and in the wage-rental ratio from the 1880s (Spree 1977 and Berghahn 1994), which is in turn related to the provision of industrial education in the second half of the 19th century.

The link between the expansion of education and the reduction in inequality is present in the US as well. Wealth inequality in the US, which increased gradually from colonial times until the second half of the 19th century, reversed its course at the turn of the century and maintained its declining pattern during the first half of the 20th century (Lindert and Williamson 1976). As argued by Goldin (2001), the emergence of the "new economy" in the early 20th century increased the demand for educated workers. The creation of publicly funded mass modern secondary schools from 1910 to 1940 provided general and practical education, contributed to workers productivity and opened the gates for college education. This expansion facilitated social and geographic mobility and generated a large decrease in inequality in economic outcomes.

C. Independence of Political Reforms in the 19th Century

The 19th century was marked by significant political reforms along with the described education reforms. One could therefore challenge the significance of the industrial motive for education reform, suggesting that political reforms during the 19th century shifted the balance of power towards the working class and enabled workers to implement education reforms against the interest of the industrial elite, has no empirical support. However, political reforms that took place in the 19th century had no apparent effect on education reforms over this period, strengthening the hypothesis that indeed industrial development, and the increasing demand for human capital, were the trigger for human capital formation and the demographic transition.[47] Education reforms took place in autocratic states that did not relinquish political power throughout the 19th century, and major reforms occurred in societies in the midst of the process of democratization well before the stage at which the working class constituted the majority among the voters.

In particular, the most significant education reforms in the UK were completed before the voting majority shifted to the working class. The patterns of education and political reforms in the UK during the 19th century are depicted in Figure 2.30. The Reform Act of 1832 nearly doubled the total electorate, but nevertheless only 13% of the voting-age population were enfranchised. Artisans, the working classes, and some sections of the lower middle classes remained outside of the political system. The franchise was extended further in the Reform Acts of 1867 and 1884 and the total electorate nearly doubled in each of these episodes. However, working-class voters did not become the majority in all urban counties until 1884 (Craig 1989).

The onset of England's education reforms, and in particular, the fundamental Education Act of 1870 and its major extension in 1880 occurred prior to the political reforms of 1884 that made the working class the majority in most counties. As depicted in Figure 2.30, a trend of significant increase in primary education was established well before the extension of the franchise in the context of the 1867 and 1884 Reform Acts. In particular, the proportion of children aged 5 to 14 in primary schools increased five-fold (and surpassed 50%) over the three decades prior to the qualitative extension of the franchise in 1884 in which the working class was granted a majority in all urban counties. Furthermore, the political reforms do not appear to have any effect on the pattern of education reform. In fact, the average growth rate of education attendance from decade to decade over the period 1855 to 1920

[47] See for instance, Acemoglu and Robinson (2000), where the extension of the franchise during the 19th century is viewed as a commitment devise ensuring future income redistribution from the elite to the masses.

reaches a peak at around the Reform Act of 1884 and starts declining thereafter. It is interesting to note, however, that the abolishment of education fees in nearly all elementary schools occurs only in 1891, after the Reform Act of 1884, suggesting that the political power of the working class may have affected the distribution of education cost across the population, but consistent with the proposed thesis, the decision to educate the masses was taken independently of the political power of the working class.

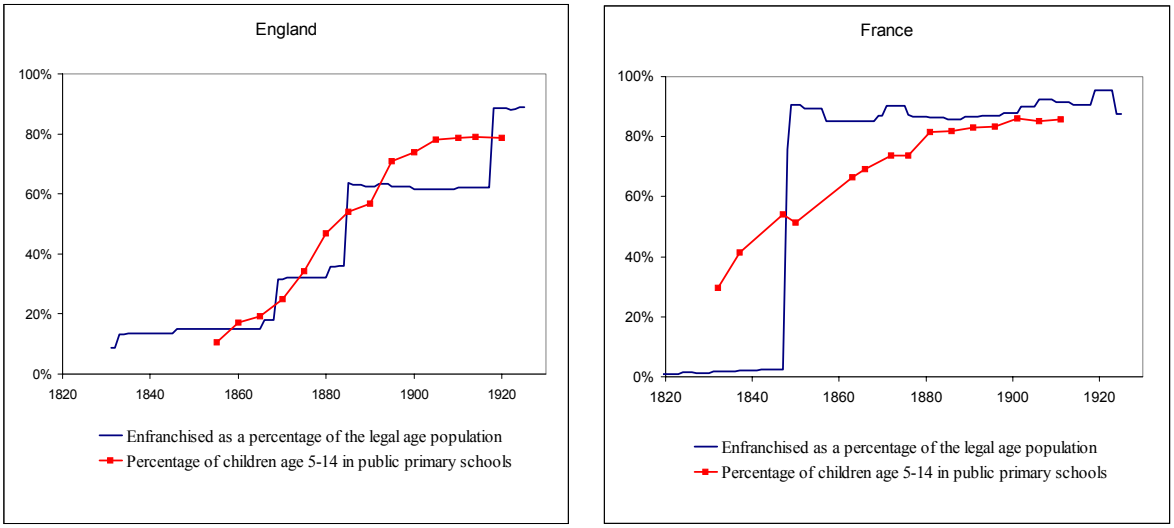


Figure 2.30. The Evolution of Voting Rights and School Enrolment
Source: Flora et al. (1983)

In France, as well, the expanding pattern of education preceded the major political reform that gave the voting majority to the working class. The patterns of education and political reforms in France during the 19th century are depicted in Figure 2.30. Prior to 1848, restrictions limited the electorate to less than 2.5% of the voting-age population. The 1848 revolution led to the introduction of nearly universal voting rights for males. Nevertheless, the proportion of children aged 5 to 14 in primary schools doubled (and exceeded 50%) over the two decades prior to the qualitative extension of the franchise in 1848 in which the working class was granted a majority among voters. Furthermore, the political reforms of 1848 do not appear to have any effect on the pattern of education expansion.

A similar pattern occurs in other European countries. Political reforms in the Netherlands did not affect the trend of education expansion and the proportion of children aged 5 to 14 in primary schools exceeded 60% well before the major political reforms of 1887 and 1897. Similarly, the trends of political and education reforms in Sweden, Italy, Norway, Prussia and Russia do not lend credence to the alternative hypothesis.

Less Developed Economies

The process of industrialization was characterized by a gradual increase in the relative importance of human capital in less developed economies as well. As depicted in Figure 2.31, educational attainment increased significantly across all less developed regions. Moreover, in line with the pattern that emerged among developed economies in the 19th century, the increase in educational attainment preceded the decline in total fertility rates. In particular, the average years of schooling in Africa increased by 44% (from 1.56 to 2.44) prior to the onset of decline in total fertility rates in 1980, as depicted in Figure 2.23, whereas the available data for Asia and Latin America demonstrates a simultaneous increase in educational attainment and a decline in fertility.

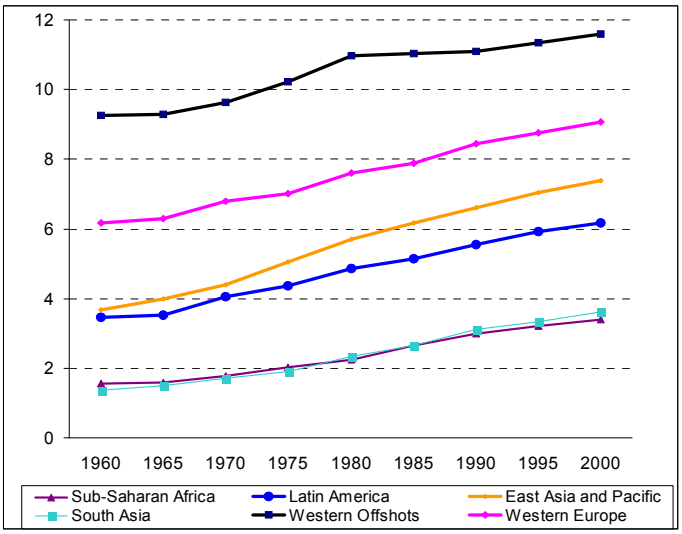


Figure 2.31. The Evolution of Average Years of Education: 1960-2000
Source: Barro and Lee (2000)

2.3.4 International Trade and Industrialization

The process of industrialization in developed economies was enhanced by the expansion of international trade. During the 19th century, North-South trade, as well as North-North trade, expanded significantly due to a rapid industrialization in Northwest Europe as well as the reduction of trade barriers and transportation costs and the benefits of the gold standard. The ratio of world trade to output was about 2% in 1800, but then it rose to 10% in 1870, to 17% in 1900 and 21% in 1913 (Estavadeordal, Frantz, and Taylor 2002). While much of this trade occurred between industrial economies a significant proportion was between industrial and non-industrial economies. As shown in Table 2.1, before 1900 nearly 50% of manufactured exports were to non-European and non-North American economies. By the end of 19th century a clear pattern of specialization emerged. The UK and Northwest Europe were net importers of primary products and net exporters of manufactured goods, whereas the exports of Asia, Oceania, Latin America and Africa were overwhelmingly composed of primary products. (Findlay and O'Rourke 2001).

Table 2.1. Regional Shares of World Trade in Manufactures
Source: Yates (1959)

	1876-1880		1896-1900		1913	
	Exports	Imports	Exports	Imports	Exports	Imports
U.K. and Ireland	37.8%	9.1%	31.5%	10.4%	25.3%	8.2%
Northwest Europe	47.1%	18.1%	45.8%	20.3%	47.9%	24.4%
Other Europe	9.2%	13.3%	10.3%	12.2%	8.3%	15.4%
U.S. and Canada	4.4%	7.7%	7.4%	9.6%	10.6%	12.1%
Rest of the World	1.5%	51.8%	5.0%	47.5%	7.9%	39.9%

Atlantic trade as well as trade with Asia, in an era of colonialism, had a major effects on European growth starting in the late 16th century (Acemoglu et al. 2003). Furthermore, later expansion of international trade contributed further to the process of industrialization in the UK and Europe (O'Rourke and Williamson 1999). For the UK, the proportion of foreign trade to national income grew from about 10% in the 1780s to about 26% over the years 1837-45, and 51.5% in the time period 1909-13 (Kuznets 1967). Other European economies experienced a similar pattern as well. The proportion of foreign trade to national income on the eve of World War I was 53.7% in France, 38.3% in Germany , 33.8% in Italy, and 40.4% in Sweden (Kuznets 1967, Table 4). Furthermore, export was critical for the viability of some industries, especially the cotton industry, where 70% of the UK output was exported in the 1870's. The quantitative study of Stokey (2001) suggests that trade was instrumental for the increased share of manufacturing in total output in the UK, as well as for the significant rise in real wages. Thus while it appears that technological advances could have spawned the Industrial Revolution without an expansion of international trade, the growth in exports increased the pace of industrialization and the growth rate of output per capita.[48]

2.4 The Great Divergence

The differential timing of the take-off from stagnation to growth across countries and the corresponding variations in the timing of the demographic transition led to a great divergence in income per capita as well as population growth.

The last two centuries have witnessed dramatic changes in the distribution of income and population across the globe. Some regions have excelled in the growth of income per capita, while other regions have been dominant in population growth. Inequality in the world economy was negligible till the 19th century. The ratio of GDP per capita between the richest region and the poorest region in the world was only 1.1:1 in the year 1000, a 2:1 in the year 1500 and 3:1 in the year 1820. As depicted in Figure 2.32, there has been a 'Great Divergence' in income per capita among countries and regions in the past two centuries. In particular, the ratio of GDP per capita between the richest region (Western offshoots) and the poorest region (Africa) has widened considerably from a modest 3:1 ratio in 1820, to a 5:1 ratio in 1870, a 9:1 ratio in 1913, a 15:1 in 1950, and a huge 18:1 ratio in 2001.

[48] Pomeranz (2000), provides historical evidence for the vital role of trade in the 'take off' of the European economies. He argues that technological and development differences between Europe and Asia were minor around 1750, but the discovery of the New World enabled Europe, via Atlantic trade, to overcome 'land constraints' and to take-off technologically.

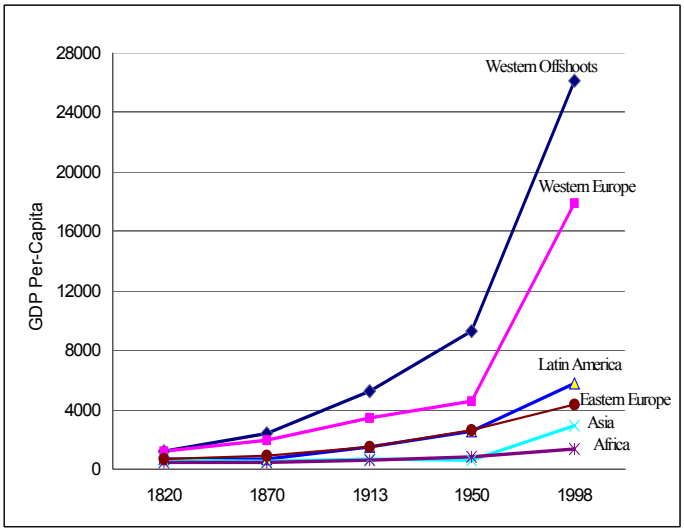


Figure 2.32. The Great Divergence
Source: Maddison (2001)

An equally momentous transformation occurred in the distribution of world population across regions, as depicted in Figure 2.33. The earlier take-off of Western European countries increased the amount of resources that could be devoted for the increase in family size, permitting a 16% increase in the share of their population in the world economy within a 50 year period (from 12.8% in 1820 to 14.8% in 1870). However, the early onset in the Western European demographic transition and the long delay in the demographic transition of less developed regions well into the 2nd half of the twentieth century led to a 55% decline in the share of Western European population in the world, from 14.8% in 1870 to 6.6% in 1998. In contrast, the prolongation of Post-Malthusian period among less developed regions in association with the delay in their demographic transition well into the second half of 20th century, channeled their increased resources towards a significant increase in their population. Africa's share of world population increased 84%, from 7% in 1913 to 12.9% in 1998, Asia's share of world population increased 11% from 51.7% in 1913 to 57.4% in 1998, and Latin American countries increased their share in world population from 2% in 1820 to 8.6% in 1998.

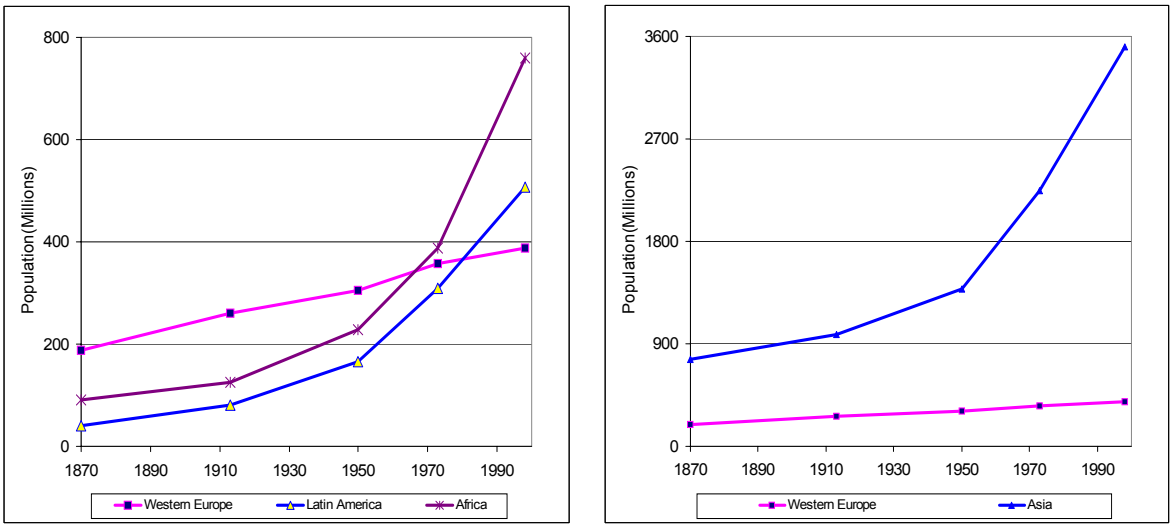


Figure 2.33. Divergence in Regional Populations

Thus, while the ratio of income per capita in Western Europe to that in Asia has tripled in the last two centuries, the ratio of Asian to European population has doubled.[49]

The divergence that has been witnessed in the last two centuries has been maintained across countries in the last decades as well (e.g., Jones 1997 and Pritchett 1997). Interestingly, however, Sala-i-Martin (2002) has shown that the phenomenon has not been maintained in recent decades across people in the world, (i.e., when national boundaries are removed).

3 The Fundamental Challenges

The discovery of a unified theory of economic growth that could account for the intricate process of development in the last thousands of years is one of the most significant research challenges facing researchers in the field of growth and development. A unified theory would unveil the underlying micro-foundations that are consistent with entire the process of economic development, enhancing the confidence in the viability of growth theory, its predictions and policy implications, while improving the understanding of the driving forces that led to the recent transition from stagnation to growth and the Great Divergence. Moreover, a comprehensive understanding of the hurdles faced by less developed economies in reaching a state of sustained economic growth would be futile unless the forces that initiated the transition of the currently developed economies into a state of sustained economic growth would be identified and modified to account for the differences in the structure of less developed economies in an interdependent world.

The evidence presented in section 2 suggests that the preoccupation of growth theory with the empirical regularities that have characterized the growth process of developed economies in the past century and of less developed economies in the last few decades, has become harder to justify from a scientific viewpoint. Could we justify the use of selective observations about the recent course of the growth process and its principal causes in the formulating exogenous and endogenous neoclassical growth models? Could we be confident about the predictions of a theory that is not based on micro-foundations that match the major characteristics of the entire growth process? The evolution of theories in older scientific disciplines suggests that theories that are founded on the basis of a subset of the existing observations are fragile and non-durable.

3.1 Mysteries of the Growth Process

The underlying determinants of the stunning recent escape from the Malthusian trap have been shaded in mystery and their significance for the understanding of the contemporary growth process has been explored only very recently. What are the major economic forces that led to the epoch of Malthusian stagnation that had characterized most of human history? What is the origin of the sudden spurt in growth rates of output per capita and population that occurred in the course of the take-off from stagnation to growth? Why had episodes of technological progress in the pre-industrialization era failed to generate sustained economic growth? What was the source of the dramatic reversal in the positive relationship between income per capita and population that existed throughout most human history? What are the main forces that prompted the demographic transition? Would the transition to a state of sustained economic growth be feasible without the demographic transition? Are there

[49] Over the period 1820-1998, the ratio between income per capita in Western Europe and Asia (excluding Japan) grew 2.9 times, whereas the ratio between the Asian population (excluding Japan) and the Western European population grew 1.7 times [Maddison, 2001].

underlying unified behavioral and technological structures that can account for these distinct phases of development simultaneously and what are their implications for the contemporary growth process?

The mind-boggling phenomenon of the Great Divergence in income per capita across regions of the world in the past two centuries, that accompanied the take-off from an epoch of stagnation to a state of sustained economic growth, presents additional unresolved mysteries about the growth process. What accounts for the sudden take-off from stagnation to growth in some countries in the world and the persistent stagnation in others? Why has the positive link between income per capita and population growth reversed its course in some economies but not in others? Why have the differences in income per capita across countries increased so markedly in the last two centuries? Has the pace of transition to sustained economic growth in advanced economies adversely affected the process of development in less-developed economies?

The transitions from a Malthusian epoch to a state of sustained economic growth and the emergence of the Great Divergence have shaped the current structure of the world economy. Nevertheless, neoclassical growth models abstracted from these the significant aspects of the growth process. In recent years, however, it has been increasingly recognized that the understanding of the contemporary growth process would be fragile and incomplete unless growth theory would be based on proper micro-foundations that reflect the growth process in its entirety.

3.2 The Incompatibility of Non-Unified Growth Theories

Existing (non-unified) growth models are unable to capture the growth process throughout human history. Malthusian models capture the growth process during the Malthusian epoch but are incompatible with the transition to the Modern Growth Regime. Neoclassical growth models (with endogenous or exogenous technological change), in contrast, are compatiable with the growth process of the *developed* economies during the Modern Growth Regime, but fail to capture the evolution of economies during the Malthusian epoch, the origin of the take-off from the Malthusian epoch into the Post-Malthusian Regime, and the sources of the demographic transition and the emergence of the modern growth regime. Moreover, the failure of non-unified growth models in identifying the underlying factors that led to the transition from stagnation to growth limits their applicability for the contemporary growth process of the less developed economies and thereby for the current evolution of the world income distribution.

3.2.1 Malthusian and Post-Malthusian Theories

The Malthusian Theory

The Malthusian theory, as was outlined initially by Malthus (1798), captures the main attributes of the epoch of Malthusian stagnation that had characterized most of human existence, but is utterly inconsistent with the prime characteristics of the modern growth regime.[50]

The theory suggests that the stagnation in the evolution of income per capita over this epoch reflected the counterbalancing effect of population growth on the expansion of resources, in an environment characterized by diminishing returns to labor. The expansion of resources, according to Malthus, would lead to an increase in population growth, reflecting the natural result of "passion between the sexes".[51] In contrast, when population size would grow beyond the capacity of the available resources,

[50]The Malthusian theory was formalized recently. Kremer (1993) models a reduced-form interaction between population and technology along a Malthusian equilibrium, and Lucas (2002) presents a Malthusian model in which households optimize over fertility and consumption, labor is subjected to diminishing returns due to the presence of a fixed quantity of land, and the Malthusian level of income per capita is determined endogenously.

[51]As argued by Malthus (1798), *"The passion between the sexes has appeared in every age to be so nearly the same, that it may always be considered, in algebraic language as a given quantity."*

it would be reduced by the "preventive check" (i.e., intentional reduction of fertility) as well as by the "positive check" (i.e., the tool of nature due to malnutrition, disease, and famine).

According to the theory, periods marked by the absence of changes in the level of technology or in the availability of land, were characterized by a stable population size as well as a constant income per capita. In contrast, episodes of technological progress, land expansion, and favorable climatic conditions, brought about temporary gains in income per capita, triggering an increase in the size of the population which led ultimately to a decline in income per capita to its long-run level. The theory proposes therefore that variations in population density across countries during the Malthusian epoch reflected primarily cross-country differences in technologies and land productivity. Due to the positive adjustment of population to an increase in income per capita, differences in technologies or in land productivity across countries resulted in variations in population density rather than in the standard of living.

The Malthusian theory generates predictions that are largely consistent with the characteristics of economies during the Malthusian epoch, as described in Section 2.1: (a) Technological progress or resource expansion would lead to a larger population, without altering the level of income in the long run. (b) Income per capita would fluctuate during the Malthusian epoch around a constant level. (c) Technologically superior countries would have eventually denser populations but their standard of living in the long run would not reflect the degree of their technological advancement. These predictions, however, are irremediably inconsistent with the relationship between income per capita and population that has existed in the post-demographic transition era as well as with the state of sustained economic growth that had characterized the Modern Growth Regime.

Unified theories of economic growth, in contrast, incorporate the main ingredients of the Malthusian economy into a broader context focusing on the interaction in this epoch between technology and the size of the population and the distribution of its characteristics, that generate the main ingredients of the Malthusian epoch but lead to an inevitable take-off to the Post Malthusian Regime.

The Post-Malthusian Theory

The Post-Malthusian theories capture the acceleration of the growth rate of income per capita and population growth that occurred in the Post-Malthusian Regime in association with the process of industrialization. They do not capture, however, the stagnation during the Malthusian epoch and the economic forces that gradually emerged in this era and brought about the take-off from the Malthusian trap. Moreover, these theories do not account for the factors that ultimately originated the demographic transition and the transition to a state of sustained economic growth.[52]

These theories suggest that the acceleration in technological progress and the associated rise in income per capita was only channeled partly towards an increase in the size of the population. Although, the Malthusian mechanism linking higher income to higher population growth continued to function, the effect of higher population on diluting resources per capita, and thus lowering income per capita, was counteracted by the acceleration in technological progress and capital accumulation, allowing income

[52]Models that are not based on Malthusian elements are unable to capture the long epoch of Malthusian stagnation in which output per capita fluctuates around a subsistence level. For instance, an interesting research by Goodfriend and McDermott (1995) demonstrates that exogenous population growth increases population density and hence generates a greater scope for the division of labor inducing the development of markets and economic growth. Their model, therefore generates a take-off from non-Malthusian stagnation to Post-Malthusian Regime in which population and output are positively related. The model lacks Malthusian elements and counter-factually it implies therefore that since the emergence of a market economy over 5000 years ago growth has been strictly positive. Moreover, it does not generate the forces that would bring about the demographic transition and ultimately sustained economic growth. In the long-run the economy remains in the Post-Malthuisan regime in which the growth of population and output are positively related. Other non-Malthusian models that abstracts from population growth and generate an acceleration of output growth along the process of industrialization include Acemoglu and Zilibotti (1997).

per capita to rise despite the offsetting effects of population growth.

Kremer (1993), in an attempt to defend the role of the scale effect in endogenous growth models, examines a reduced-form of the coevolution of population and technology in a Malthusian and Post Malthusian environment, providing evidence for the presence of a scale effect in the pre-demographic transition era. [53] Kremer's Post-Malthusian theory does not identify the factors that brought about the take-off from the Malthusian trap, as well as the driving forces behind the demographic transition and the transition to a state of sustained economic growth.

Unified theories capture the main characteristics of the Post-Malthusian Regime, and generate, in contrast, the endogenous driving forces that brought about the take-off from the Malthusian epoch into this regime and ultimately enabled the economy to experience a demographic transition and to reside in a state of sustained economic growth.

3.2.2 Theories of Modern Economic Growth

Exogenous growth models (e.g. Solow 1956) that have focused primarily on the role of factor accumulation in the growth process, as well as endogenous growth models (e.g., Romer 1990, Grossman and Helpman 1991, and Aghion and Howitt 1992) that have devoted their attention to the role of endogenous technological progress in the process of development, were designed to capture the main characteristics of the Modern Growth Regime. These models, however, are inconsistent with the pattern of development that had characterized economies over most of human existence. They do not account for Malthusian epoch the economic factors that brought about the take-off from the Malthusian regime into the Post-Malthusian Regime, and the forces that brought about the demographic transition and ultimately the state of sustained economic growth.[54]

Modern non-unified growth theory has not developed the research methodology that would enable researchers to shed light on the principal factors that would enable less developed economies that are in a state of Malthusian stagnation, or in a post-Malthusian regime to take-off to a state of sustained economic growth. Moreover, most endogenous and exogenous growth models are inconsistent with the changes in the demographic regime along the process of development.[55] With few exceptions non-unified growth models do not generate the hump-shaped relationship between income per capita and population growth in the process of development. Most growth models with endogenous population have been oriented toward the modern regime, capturing the recent negative relationship between population growth and income per capita, but failing to capture the positive effect of income per capita on population growth that had characterized most human existence and the economic factors that triggered the demographic transition.[56]

[53] Komlos and Artzrouni (1990) simulates an escape from a Malthusian trap based on the Malthusian and Boserupian interaction between population and technology.

[54] Non-unified growth models are inconsistent with the process of development in the Malthusian epoch. Moreover, as long as the neoclassical production structure of non-decreasing returns to scale is maintained, they could not be modified to account for the Malthusian epoch by the incorporation of endogenous population growth. Suppose that the optimal growth model would be augmented to account for endogenous population. Suppose further that the parameters of the model would be chosen so as to assure that the level of income per capita would reflect the level that existed during the Malthusian epoch and population growth will be near replacement level as was the case during this era. This equilibrium would not possess the prime characteristic of a Malthusian equilibrium. Namely, technological progress would raise income per capita permanently since adjustments in population growth would not offset this rise of income (as long as the return to labor is characterized by non-diminishing returns to scale).

[55] In fact, most endogenous growth models that focus exclusively on the modern growth regime are inconsistent with the demographic structure within this regime, predicting a positive effect of population growth on (the growth rate of) income per capita. A notable exception is Dalgaard and Kreiner (2001).

[56] Earlier papers that captures aspects of the cross-section relationship between income per capita and fertility includes Ben Zion and Razin (1975), Barro and Becker, (1989) and Becker, Murphy and Tamura (1990), and recent ones include Deopke and De la Croix (2003) and Moav (2005).

3.3 Theories of the Demographic Transition and Their Empirical Validity

The theories of the demographic transition attempt to capture the determinants of the significant reduction in fertility rates and population growth that characterized the world in the past century, following the unprecedented increase in population growth during the Post-Malthusian regime, enabling economies to convert a increasing share of the benefits of factor accumulation and technological progress into growth of output per capita.

There are several factors that could have theoretically triggered a demographic transition. The simultaneity of the demographic transition across Western European countries suggests that a common cause may have originated the various transitions. Was it an outcome of a simultaneous decline in mortality rates across Western European countries? Was it associated with a nearly simultaneous rise in income across Western European countries? An outcome of the rise in the relative wages of women in the second phase of the Industrial Revolution? Or was it a consequence of the universal rise in the demand for education and the associated decline in child labor in the second phase of the Industrial Revolution? Historical evidence suggests that demographers' explanation of the demographic transition - the decline in mortality - is highly implausible. Moreover, Becker's emphasis on the role of rising income in the demographic transition is inconsistent with the evidence. Empirical evidence suggests that the rise in the demand for human capital is the most significant force behind the demographic transition and it is therefore a critical building block in existing unified theories.

3.3.1 The Decline in Infant and Child Mortality

The decline in infant and child mortality rates that preceded the decline in fertility rates in most countries in the world, with the notable exceptions of France and the US, has been demographers' favorite explanation for the onset of the decline in fertility in the course of the demographic transition.[57] Nevertheless, it appears that this simplistic viewpoint is based on weak theoretical reasonings and is inconsistent with historical evidence.

Existing theories suggest that parents generate utility from the number (and possibly the quality) of their surviving offspring. A decline in mortality rates, therefore, would be expected to lead to a corresponding reduction in total fertility rates, but not necessarily to a reduction in the number of children reaching adulthood. While it is highly plausible that mortality rates were among the factors that affected the level of total fertility rates along human history, historical evidence does not lend credence to the argument that the decline in mortality rates accounts for the *reversal* of the positive historical trend between income and fertility. A careful examination of the various factors that affect fertility rates (i.e., mortality rate, income level, the return to investment in child quality, and the gender wage gap) reveals that this argument is inconsistent with the evidence.

The decline in mortality rates does not appear to be the trigger for the decline in fertility in Western Europe. As demonstrated in Figures 2.23 and 2.24, the mortality decline in Western Europe started nearly a century prior to the decline in fertility and it was associated initially with increasing fertility rates in some countries and non-decreasing fertility rates in other countries

In particular, as demonstrated in Figure 3.1, the decline in mortality started in England in the 1730s and was accompanied by a steady increase in fertility rates until 1820. The significant rise in income per capita in the Post-Malthusian Regime increased the desirable number of surviving offspring and thus, despite the decline in mortality rates, fertility increased significantly so as to reach this higher

[57] The effect of the decline in mortality rates on the prolongation of productive life and thus on the return to human capital is discussed in section 3.3.3.

desirable level of surviving offspring.[58] As depicted in Figure 3.1, the decline in fertility during the demographic transition occurred in a period in which this pattern of increased income per capita (and its potential positive effect on fertility) was intensified, while the pattern of declining mortality (and its adverse effect on fertility) maintained the trend that existed in the 140 years that preceded the demographic transition, suggesting that in the absence of the intervention of a third factor fertility would have risen further.[59] Thus, the reversal in the fertility patterns in England as well as other Western European countries in the 1870s suggests that it is very likely that the demographic transition was prompted by a different universal force.[60]

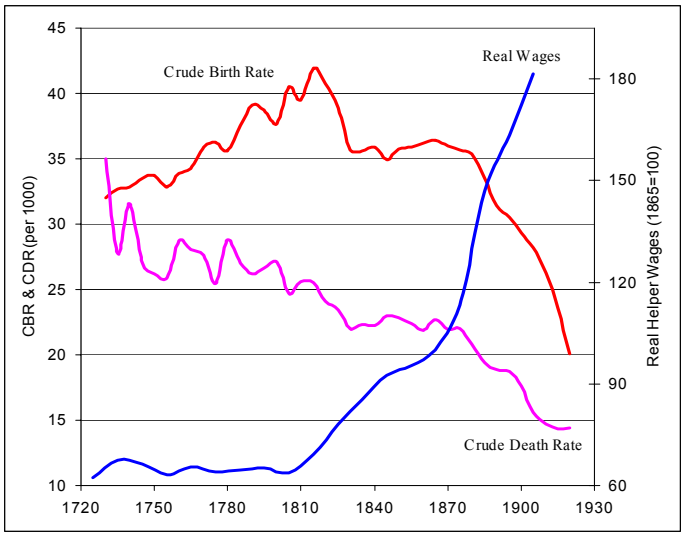


Figure 3.1. Fertility, Mortality and Real Wages, England 1730-1920
Source: Wrigley and Schofield (1981), Clark (2002), and Andorka (1978)

Furthermore, most relevant from an economic point of view is the cause of the reduction in net fertility (i.e. the number of children reaching adulthood). The decline in the number of surviving offspring that was observed during the demographic transition (e.g., Figure 2.22) is unlikely to follow from mortality decline. Mortality decline would lead to a reduction in the number of surviving offspring if the following set implausible of conditions would be met:[61] (i) there exists a significant precautionary demand for children, i.e., individuals are significantly risk averse with respect to their expected number of surviving offspring and they hold a buffer stock of children in a high mortality environment (highly improbable from an evolutionary perspective), (ii) risk aversion with respect to consumption is smaller than risk aversion with respect to fertility (evolutionary theory would suggest the opposite), (iii) sequential fertility (i.e., replacement of non-surviving children) is modest,[62] and (iv) parental resources

[58]The argument is valid even if fertility rates would have remained unchanged over this period.

[59]One could argue that the decline in mortality was not internalized into the decision of household who had difficulties separating temporary decline from a permanent one. This argument is highly implausible given the fact that mortality declined monotonically for nearly 140 years prior to the demographic transition. It is inconceivable that six generations of household did not update information about mortality rates in their immediate surrounding, while keeping the collective memories about mortality rates two centuries earlier.

[60]The mortality channel is inconsistent with additional evidence: the decline in fertility started in France and the US prior to the decline in mortality rates

[61]In particular, the theoretical analysis of Kalemli-Ozcan (2002) generates a reduction in net fertility in reaction to a decline in mortality assuming (implicitly) that all these implausible conditions are satisfied. Eckstein et al. (1999) argue in their structural quantitative analysis of the demographic transition in Sweden, that mortality decline played a role in the demographic transition. Their underlying theoretical structure, however, requires conditions (iii) and (iv) as well as specific interactions between mortality, wages, and the return to human capital.

[62]Doepke (2005) shows that regardless of the degree of risk aversion, the feasibility of sequential fertility is sufficient to preclude the decline in net fertility in reaction to a decline in mortality.

saved from the reduction in the number of children that do not survive to adulthood do not lead to a rise in fertility.[63]

A quantitative analysis by Doepke (2005) supports the viewpoint that a decline in infant mortality rates was not the trigger for the decline in net fertility during the demographic transition. Utilizing the mortality and fertility data from England in the time period 1861–1951, he shows that the decline in child mortality in this time period should have resulted in an rise in net fertility rates, in contrast to the evidence, suggesting that other factors generated the demographic transition. Similar conclusions about the insignificance of the mortality decline in the decline in fertility during the demographic transition is reached in the quantitative analysis of Fernandez-Vilaverde (2003).

3.3.2 The Rise in the Level of Income Per Capita

The rise in income per capita prior to the demographic transition has led some researchers to argue that the demographic transition was triggered by the rise in income per capita and its asymmetric effects on the income of households on the one hand and the opportunity cost of raising children on the other hand.

Becker (1981) advanced the argument that the decline in fertility in the course of the demographic transition is a by-product of the rise in income per capita that preceded the demographic transition. He argues that the rise in income induced a fertility decline because the positive income effect on fertility that was generated by the rise in wages was dominated by the negative substitution effect that was brought about by the rising opportunity cost of children. Similarly, Becker and Lewis (1973) argue that the income elasticity with respect to child quality is greater than that with respect to child quantity, and hence a rise in income led to a decline in fertility along with a rise in the investment in each child.

This theory, however, is counter-factual. It suggests that the timing of the demographic transition across countries in similar stages of development would reflect differences in income per capita. However, remarkably, as depicted in Figure 2.22, the decline in fertility occurred in the same decade across Western European countries that differed significantly in their income per capita. In 1870, on the eve of the demographic transition, England was the richest country in the world, with a GDP per capita of \$3191.[64] In contrast, Germany that experienced the decline in fertility in the same years as England, had in 1870 a GDP per capita of only \$1821 (i.e., 57% of that of England). Sweden's GDP per capita of \$1664 in 1870 was 48% of that of England, and Finland's GDP per capita of \$1140 in 1870 was only 36% of that of England, but their demographic transitions occurred in the same decade as well.

The simultaneity of the demographic transition across Western European countries that differed significantly in their income per capita suggests that the high level of income that was reached by Western Europeans countries in the Post-Malthusian regime had a very limited role in the demographic transition.[65] Furthermore, a quantitative analysis of the demographic transition in England, conducted by Fernandez-Vilaverde (2003), demonstrates that Becker's theory is counter-factual. In contrast to Becker's theory, the calibration suggests that a rise in income would have resulted in an increase in fertility rates, rather than in the observed decline in fertility.

Interestingly, and consistent with subsequent theories that underlined the critical role of technological progress in the demographic transition, despite the large differences in the *levels* of income

[63] An additional force that operates against the decline in the number of surviving offspring is the physiological constraint on the feasible number of birth per woman. If this constraint is binding for some households in a high mortality regime, a reduction in mortality would operate towards an increase the number of surviving offspring.

[64] Source: Maddison (2001). GDP per capita is measured in 1990 international dollars.

[65] Furthermore, cross-section evidence within countries suggest that the elasticity of the number of surviving offspring with respect to wage income was positive prior to the demographic transition (e.g., Clark (2003)), in contrast to Becker's argument that would require a hump-shaped relationship.

per capita across European countries that experiences the demographic transition in the same time period, the *growth rates* of income per capita of these European countries were rather similar during their demographic transition, ranging from 1.9% per year over the period 1870-1913 in the UK, 2.12% in Norway, 2.17% in Sweden, and 2.87% in Germany.

3.3.3 The Rise in the Demand for Human Capital

The gradual rise in the demand for human capital in the second phase of the Industrial Revolution as well as in the process of industrialization of less developed economies, as documented in section 2.3.3, and its close association with the timing of the demographic transitions, has led researchers to argue that the increasing role of human capital in the production process (rather than the rise in income) induced households to increase their investment in the human capital of their offspring, ultimately leading to the onset of the demographic transition.

Galor and Weil (1999, 2000), argue that the acceleration in the rate of technological progress increased gradually the demand for human capital in the second phase of the Industrial Revolution, inducing parents to invest in the human capital of their offspring. The increase in the rate of technological progress and the associated increase in the demand for human capital brought about two effects on population growth. On the one hand, improved technology eased households' budget constraints and provided more resources for quality as well as quantity of children. On the other hand, it induced a reallocation of these increased resources toward child quality. In the early stages of the transition from the Malthusian regime, the effect of technological progress on parental income dominated, and the population growth rate as well as the average quality increased. Ultimately, further increases in the rate of technological progress that were stimulated by human capital accumulation induced a reduction in fertility rates, generating a demographic transition in which the rate of population growth declined along with an increase in the average level of education. Thus, consistent with historical evidence, the theory suggests that prior to the demographic transition, population growth increased along with investment in human capital, whereas the demographic transition brought about a decline in population growth along with a further increase in human capital formation.[66]

Moreover, Galor and Weil's theory suggests that a universal rise in the demand for human capital in Western Europe (as documented in section 2.3.3) followed by a rapid growth rates of output per-capita generated the observed simultaneous onset of the demographic transition across Western European countries that differed significantly in their levels of income per capita. The simultaneous increase in educational attainment across Western European countries in the second phase of the Industrial Revolution appears as a plausible common cause that brought about the demographic transition. The rise in the demand for human capital in the second phase of the Industrial Revolution in some Western European countries (as documented in section 2.3.3) and the expectations for an imminent increase in the demand for human capital in the other Western European countries led to a significant increase in the investment in children's education and therefore to a decline in fertility.

In particular, as depicted in Figure 3.2, the demographic transition in England was associated with a significant increase in the investment in child quality as reflected by years of schooling. Quantitative evidence provided by Doepke (2004) suggests that indeed, educational policy played an important role in the demographic transition in England.

[66] Quantitative evidence provided by Greenwood and Seshadri (2002) is supportive of the role of the rise in the demand for skilled labor in the demographic transition in the US. They demonstrate that technological progress in an industrial, skilled-intensive, sector that is larger than that in an unskilled-intensive, agricultural sector matches the data on the US demographic transition.

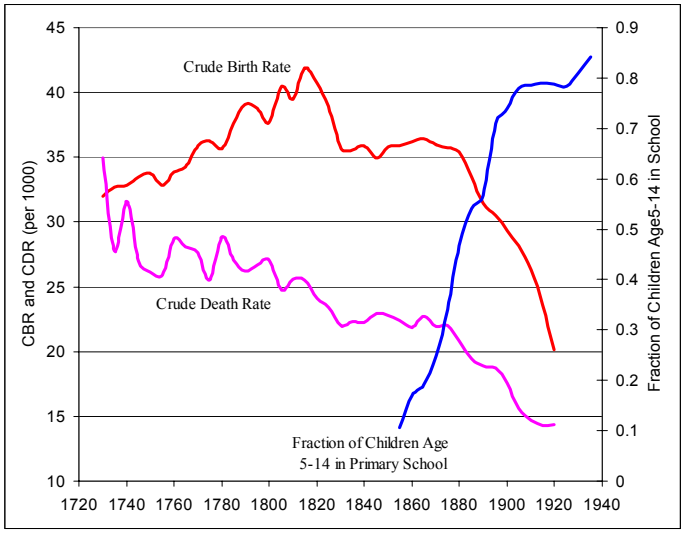


Figure 3.2. Investment in Human Capital and the Demographic Transition, England, 1730-1935
Sources: Flora et al. (1983) and Wrigley and Schofield (1983)

Reinforcing Mechanisms

The Decline in Child Labor

The effect of the rise in the demand for human capital on the reduction in the desirable number of surviving offspring was magnified via its adverse effect on child labor. It gradually increased the wage differential between parental labor and child labor inducing parents to reduce the number of their children and to further invest in their quality (Hazan and Berdugo 2002).[67] Moreover, the rise in the importance of human capital in the production process induced industrialists to support education reforms (Galor and Moav 2004) and thus laws that abolish child labor (Doepke and Zilibotti 2003), inducing a reduction in child labor and thus in fertility. Doepke (2004) provides quantitative evidence that suggests that indeed, child labor law, and to a lesser extent educational policy, played an important role in the demographic transition.

The Rise in Life Expectancy

The impact of the increase in the demand for human capital on the decline in the desirable number of surviving offspring may have been reinforced by the rise in life expectancy.

The decline in mortality rates in developed countries since the 18th century, as depicted in Figure 2.24, and the recent decline in mortality rates in less developed countries, as depicted in Figure 2.25, corresponded to a gradual increase in life expectancy. As depicted in Figure 2.26, life expectancy in England increased at a stable pace from 32 years in the 1720s to about 41 years in the 1870. This pace of the rise in life expectancy increased and life expectancy reached 50 years in 1900, 69 years in 1950 and 77 years in 1999. Similarly, life expectancy increased in Western Europe during the 19th century from 36 in 1820 to 46 in 1900, 67 in 1950 and 78 in 1999. In less developed economies life expectancy increased markedly in the 20th century, as depicted in Figure 2.27.

[67] Hazan and Berdugo (2002)'s hypothesis is consistent with existing historical evidence. For instance, Horrell and Humphries (1995) suggest, based on data from the United Kingdom, that the earnings of children age 10-14 as a percentage of father's earning declined from the period 1817-1839 to the period 1840-1872 by nearly 50% if the father was employed in a factory. Interestingly, the effect is significantly more pronounced if the father was employed in skilled occupations (i.e., high wage agriculture) rather than low skilled occupations (i.e., mining), reflecting the rise in the relative demand for skilled workers and its effect on the decline in the relative wages of children.

Despite the gradual rise in life expectancy prior to the demographic transition investment in human capital was rather insignificant as long as a technological demand for human capital had not emerged. In particular, the increase in life expectancy in England occurred 150 years prior to the demographic transition and may have resulted in a gradual increase in literacy rates, but not at a sufficient level to induce a reduction in fertility. Similarly, the rise in life expectancy in less developed regions in the first half of the 20th century has not generated a significant increase in education and a demographic transition.

In light of the technologically based rise in the demand for human capital in the second phase of the Industrial Revolution, as documented in section 2.3.3, the rise in the expected length of productive life has increased the potential rate of return to investments in children's human capital, and thus re-enforced and complemented the inducement for investment in education and the associated reduction in fertility rates.[68]

Changes in Marriage Institutions.

The effect of the rise in the demand for human capital on the desirable quality of children, and thus on the decline in fertility was reinforced by changes in marriage institutions. Gould, Moav and Simhon (2003) suggest that the rise in the demand for human capital increased the demand for quality women who have a comparative advantage in raising quality children, increasing the cost of marriage. Polygamy therefore became less affordable, inducing the transition from polygamy to monogamy, and reinforcing the decline in fertility. Edlund and Lagerlof (2002) suggest that love marriage, as opposed to arranged marriage, redirected the payment for the bride from the parent to the couple, promoting investment and human capital accumulation and thus reinforcing the decline in fertility.

Natural Selection and the Evolution of Preference for Offspring's Quality

The impact of the increase in the demand for human capital on the decline in the desirable number of surviving offspring may have been magnified by cultural or genetic evolution in the attitude of individuals toward child quality. An evolutionary change in the attitude of individuals towards human capital could have generated a swift response to the increase in demand for human capital, generating a decline in fertility along with an increase in human capital formation.

Human beings, like other species, confront the basic trade-off between offspring's quality and quantity in their implicit Darwinian survival strategies. Preference for child quantity as well as for child quality reflects the well-known variety in the quantity-quality survival strategies (or in the K and r strategies) that exists in nature (e.g., MacArthur and Wilson 1967) and the allocation of resources between offspring quantity and quality is subjected to evolutionary changes (Lack 1954).

Galor and Moav (2002) propose that during the epoch of Malthusian stagnation that characterized most of human existence, individuals with a higher valuation for offspring quality gained an evolutionary advantage and their representation in the population gradually increased. The agricultural revolution facilitated the division of labor and fostered trade relationships across individuals and communities,

[68]This mechanism was outlined by Galor and Weil (1999) and was examined in different settings by Erlich and Lui (1991), and Hazan and Zoabi (2004). It should be noted, however, that as argued by Moav (2005), the rise in the potential return to investment in child quality due to the prolongation of the productive life is not as straightforward as it may appear. It requires that the prolongation of life would affect the return to quality more than the return to quantity. For example, if parents derive utility from the aggregate wage income of their children, prolongation of life would increase the return to quantity and quality symmetrically. Hence, additional mechanism that would increase the relative complementarity between life expectancy and human capital would be needed to assure the rise in the return to human capital. For instance, Hazan and Zoabi (2004) assume that an increase in life expectancy, and thus the health of students, enhances the production process of human capital and thus increases the relative return to child quality. Alternatively, Moav (2005) argues that an increase in life expectancy, while having no effect on parental choice between quality and quantity, induce the offspring to increase their own human capital bringing about lower fertility rates in the next generation due to the comparative advantage of educated parents in educating their children.

enhancing the complexity of human interaction and raising the return to human capital. Moreover, the evolution of the human brain in the transition to Homo sapiens and the complementarity between brain capacity and the reward for human capital has increased the evolutionary optimal investment in the quality of offspring. The distribution of valuation for quality lagged behind the evolutionary optimal level and individuals with traits of higher valuation for their offspring's quality generated higher income and, in the Malthusian epoch when income was positively associated with aggregate resources allocated to child rearing, a larger number of offspring. Thus, the trait of higher valuation for quality gained the evolutionary advantage, and the Malthusian pressure gradually increased the representation in the population of individuals whose preferences were biased towards child quality.

This evolutionary process was reinforced by its interaction with economic forces. As the fraction of individuals with high valuation for quality increased, technological progress intensified, raising the rate of return to human capital. The increase in the rate of return to human capital along with the increase in the bias towards quality in the population reinforced the substitution towards child quality, setting the stage for a more rapid decline in fertility along with a significant increase in investment in human capital.

This mechanism is consistent with the gradual rise in literacy rates prior to the Industrial Revolution, as depicted in Figure 5.1. It suggests that the increase in the investment in human capital prior to the Industrial Revolution was a reflection of changes in the composition of preference for quality in the population that stimulated investment in human capital, prior to the increase in the demand for human capital in the second phase of the Industrial Revolution.

3.3.4 The Decline in the Gender Gap

The rise in women's relative wages in the last two centuries and its potential impact on the rise in female labor force participation and the associated decline in fertility rates have been the center of another theory of the demographic transition that generates the observed hump-shaped relationship between income per capita and population growth, as depicted in Figure 2.15.

The rise in women's relative wages along with declining fertility rates has been observed in a large number of developed and less developed economies. In particular, as depicted in Figure 3.3, this pattern is observed in the US during the period 1800-1940.

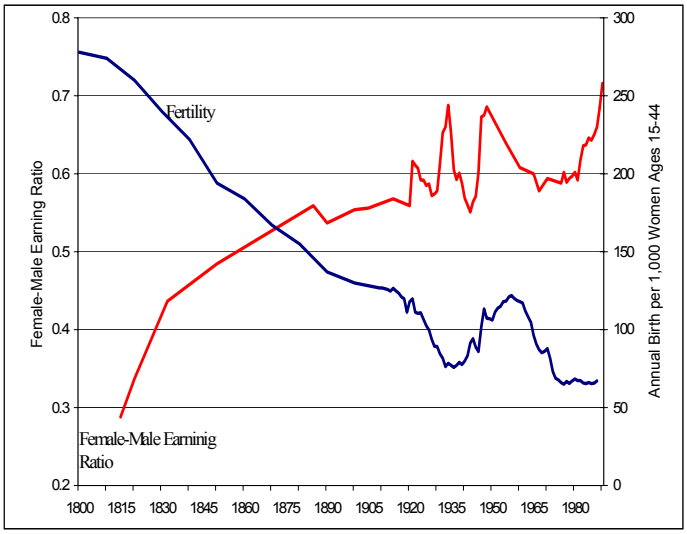


Figure 3.3. Female Relative Wages and Fertility Rates
United States 1800-1990
Source: U.S. Bureau of the Census, (1975), Goldin (1990), and Hernandez (2000)

Galor and Weil (1996) argue that technological progress and capital accumulation increased the relative wages of women in the process of industrialization. They maintain that technological progress along with physical capital accumulation complemented mental-intensive tasks rather than physical-intensive tasks and thus, in light of the comparative advantage of women in mental-intensive tasks, the demand for women's labor input gradually increased in the industrial sector, increasing the absolute wages of men and women but decreasing the gender wage gap. As long as the rise in women wages was insufficient to induce a significant increase in women's labor force participation, fertility increased due to the rise in men's wages.[69] Ultimately, however, the rise in women's relative wages was sufficient to induce a significant increase in labor force participation, generating a demographic transition.[70] Unlike the single-parent model in which an increase in income generates conflicting income and substitution effect that cancel one another if preferences are homothetic, in the two-parent household model, if most of the burden of child rearing is placed on women, a rise in women's relative wages increases the opportunity cost of raising children more than the household income, generating a pressure towards a reduction in fertility.

Moreover, the process of development in the Post-Malthusian Regime was associated with a gradual decline in the human capital gap between male and female. As depicted in Figure 3.4, literacy rates among women which were in 1840 only 76% of those among men, grew faster in the 19th century

[69]The U-shaped pattern of female labor force participation in the process of industrialization follows from the coexistence of an industrial sector and a non-modern production sector that is not fully rival with child rearing. Women's marginal product in non-modern sector was not affected by capital accumulation in the industrial sector, while women's potential wages in the modern sector increased. In the early process of industrialization, therefore, capital accumulation increased labor productivity in the industrial sector, family income increased via men's wages, while female wages, based on the production of market goods in the home sector did not change. Fertility increased due to the income effect and female labor force participation fell. Once capital accumulation and technological progress increases female relative wages sufficiently, capital accumulation raised women's relative wages, inducing a rise in female labor force participation in the industrial sector and reducing fertility.

[70]Cavalcanti and Tavares (2003) demonstrate theoretically and empirically that the decline in the gender wage gap and the increase in labor force participation increase government expenditure on public goods reducing the cost of child rearing, further enhancing the decline in fertility.

reaching men's level in 1900.

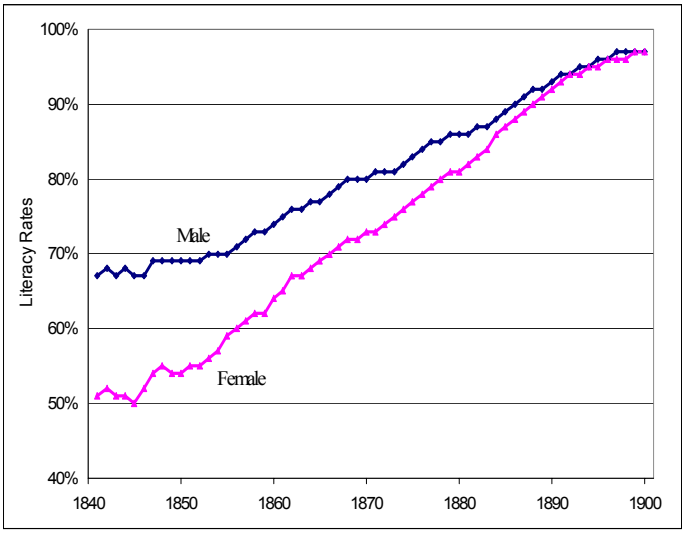


Figure 3.4. The Decline in the Human Capital Gap between Male and Female: England 1840-1900
Source: Cipolla(1969)

Lagerlof (2003b) argues that the process of development permitted a gradual improvement in the level of female education, raising the opportunity cost of children and initiating a fertility decline.

3.3.5 Other Theories

The Old-Age Security Hypothesis

The old-age security hypothesis has been proposed as an additional mechanism for the onset of the demographic transition. suggests that in the absence of capital markets that permit intertemporal lending and borrowing, children are assets that permit parents to smooth consumption over their life-time.[71] The process of development and the establishment of capital markets reduce this motivation for rearing children, contributing to the demographic transition.

Although old-age support is a plausible element that may affect the level of fertility, it appears as a minor force. The importance of the decline in the role of children as assets in the onset of the demographic transition is questionable. The rise in fertility rates prior to the demographic transition, in a period of improvements in the credit markets, raises doubts about the significance of the mechanism. Furthermore, cross-section evidence does not indicate that wealthier individuals that presumably had a better access to credit markets had a small number of surviving offspring. On the contrary, fertility rates in the pre-demographic transition era are positively related to skills, income and wealth (e.g., Lee 1997 and Clark 2003).[72]

Exogenous Shocks - Luck

Becker, Murphy and Tamura (1990) advance a theory that emphasizes the role of a major exogenous shock in triggering the demographic transition, underlying the role of luck in the determination of the relative timing of the demographic transition across nations as well as the wealth of nations.[73]. They

[71] See Neher (1971) and Caldwell (1976) for earlier studies and Boldrin and Jones (2002) for a recent quantitative analysis.

[72] Moreover, examples of species in nature in which offspring support their parent in old age are very rare.

[73] As they argue in page S13, "Many attempts to explain why some countries have had the best economic performance in the past several centuries give too little attention to accidents and good fortune".

argue that a major shock shifted economies from the low-output high-fertility steady-state equilibrium towards the high-output low-fertility steady-state equilibrium, triggering a demographic transition.

This theory suffers from critical deficiencies in the micro-structure and it generates predictions that are inconsistent with the evidence.[74] Moreover, even if economically plausible micro-foundations for the presence of multiple steady-state equilibria would be established, a major shock that would shift the economy from the basin of attraction of a high-fertility to a low-fertility steady-state equilibrium would not account for the main characteristics of the demographic transition and the transition to modern growth. In contrast to existing evidence that shows that for a large number of countries population growth rates increased prior to their decline in the course of the demographic transition and that the rise in income per capita preceded the decline in fertility,.a major shock that would shift the economy from the basin of attraction of a high-fertility to a low-fertility steady-state equilibrium would generate, counterfactually, a monotonic decline in fertility rates along with a simultaneous rise in income per capita.

4 Unified Growth Theory

The inconsistency of exogenous and endogenous neoclassical growth models with the process of development along most of human history, induced growth theorists to develop a unified theory of economic growth that would capture in a single framework the epoch of Malthusian stagnation, the era of modern growth, and the principal factors that brought about the transition between these regimes.[75]

The establishment of a unified growth theory has been a great intellectual challenge, requiring major methodological innovations in the construction of dynamical systems that would capture the complexity that has characterized the evolution of economies from a Malthusian epoch to a state of sustained economic growth. Historical evidence suggests that the take-off from the Malthusian epoch to a state of sustained economic growth, rapid as it may appear, was a gradual process and thus could not plausibly (and meaningfully) be viewed as the outcome of a major exogenous shock that shifted the economy from the basin of attraction of the Malthusian epoch into the basin of attraction of the Modern growth regime.[76] The simplest methodology for the generation of this phase transition - a major shock in an environment characterized by multiple locally stable equilibria - is therefore not applicable for the generation of the transition from stagnation to growth.

An alternative methodology, however, was rather difficult to establish since a unified growth

[74]The source of multiplicity of equilibria in their model is the implausible assumption that the return to education increases with the aggregate level of education in society. (Browing, Hansen and Heckman (1998), for instance, show that there is weak empirical evidence in favor of this assumption.) Moreover, they define erroneously the low-output, high population growth, steady-state as a Malthusian steady-state equilibrium. Their "Malthusian" steady-state, however, has none of the features of a Malthusian equilibrium. In contrast to the historical evidence about the Malthusian era, in this equilibrium (in the absence of technological change) population growth rate is not at the reproduction level. Moreover, counter-factually population growth in their "Malthusian" steady-state is *higher* than the that in the beginning of the demographic transition. Furthermore, a small positive shock to income when the economy is in the "Malthusian" steady-state initially decreases fertility in contrast to the central aspect of the Malthusian equilibrium.

[75]Growth theories that capture the evolution of population, technology, and output from stagnation to sustained economic growth have been established by Galor and Weil (1999, 2000), Galor and Moav (2002), Hansen and Prescott (2002), Jones (2001), Kogel and Prskawetz. (2001), Hazan and Berdugo (2002), Tamura (2002), Lagerlof (2003a, 2003b), Fernandez-Vilaverde (2003), Doepke (2004), as well as others. The Great Divergence and its association with the transition from stagnation to growth was explored in a unified setting by Galor and Mountford (2003).

[76]As established in section 2, and consistently with the revisionist view of the Industrial Revolution, neither the take-off of the currently developed world in the 19th century, nor the recent take-off of less developed economies provide evidence for an unprecedented shock that generated a quantum leap in income per-capita. Moreover, technological progress could not be viewed as a shock to the system. As argued by Mokyr (2002) technological progress during the Industrial Revolution was an outcome of a gradual endogenous process that took place over this time period. Moreover, technological progress in less developed economies was an outcome of a deliberate decision by entrepreneurs to adopt existing advanced technologies.

theory in which economies take-off gradually but swiftly from an epoch of a stable Malthusian stagnation would necessitate a *gradual* escape from an absorbing (stable) steady-state equilibrium, in contradiction to the essence of a stable steady-state equilibrium. Ultimately, it has become apparent that the observed gradual, rapid, and continuos phase transition would be captured by a single dynamical system, only if the set of steady-state equilibria and their stability would be altered qualitatively in the process of development. As proposed by Galor and Weil (2000), during the Malthusian epoch, the dynamical system would have to be characterized by a stable Malthusian steady-state equilibrium, but ultimately due to the evolution of latent state variables in this epoch, the Malthusian steady-state equilibrium would vanish endogenously leaving the arena to the gravitational forces of the emerging Modern Growth Regime.

Unified growth theories that permit a phase transition and do not rely on the existence of a major exogenous shock, have to be founded on a dynamical system in which economies are for an epoch in the vicinity of a temporary stable Malthusian steady-state equilibrium, but the evolution of latent state variables (i.e., the rise in a latent demand for human capital in Galor and Weil (2000) and the evolution of the distribution of genetic characteristics in Galor and Moav (2002)) ultimately generates a structural change in the dynamical system that causes this Malthusian steady state to vanish endogenously, permitting the economy to take-off and to converge to a modern growth steady-state equilibrium.

The role of the demographic transition in the transition from the Post-Malthusian Regime to the Sustained Growth Regime adds to the complexity of the desirable dynamical system. Capturing this additional transition would require the unified theory to generate endogenously a reversal in the positive Malthusian effect of income on population once the take-off occurs and to provide the reduction in fertility a special role in the transition to a state of sustained economic growth.

This section explores mechanisms that can account for the complexities of these long transitions from stagnation to growth and the emergence of the Great Divergence, focusing on the role of population, technology, income distribution and education in this intricate process. It describes several unified growth theories that encompass the transition between three distinct regimes that have characterized the process of economic development: The Malthusian Epoch, The Post-Malthusian Regime, and the Sustained Growth Regime.[77] Imposing the constraint that a unified theory would account for the entire intricate process of development in the last thousands of years is a discipline that would enhance the viability of growth theory. A unified theory of economic growth would reveal the underlying micro-foundations that are consistent with the process of economic development along the entire spectrum of human history, rather that with the last century only. It would therefore enhance the confidence in the viability of growth theory, its predictions and policy implications, while improving the understanding of the origin of the recent transition from stagnation to growth and the associated phenomenon of the great divergence.

4.1 From Stagnation to Growth

The first unified growth theory in which the endogenous evolution of population, technology, and income per capita is consistent with the process of development in the last thousands of years was advanced by Galor and Weil (2000). The theory captures the three regimes that have characterized the process of

[77] Although the emphasis is on the experience of Europe and its offshoots, since these were the areas that completed the transition from the Malthusian regime to modern growth, these theories could be modified to account for the incomplete transition of the less developed countries, integrating the drastic influence of the import of pre-existing production and health technologies.

development as well as the fundamental driving forces that generated the transition from an epoch of Malthusian stagnation to a state of sustained economic growth.

The theory proposes that in early stages of development the economy was in a stable Malthusian steady state equilibrium. Technology advanced rather slowly, and generated proportional increases in output and population. The inherent positive interaction between population and technology in this epoch, however, gradually increased the pace of technological progress and the delayed adjustment of population permitted output per capita to creep forward at a miniscule rate. The slow pace of technological progress in the Malthusian epoch provided a limited scope for human capital in the production process and parents therefore had no incentive to reallocate resources towards child quality during this era.

The Malthusian interaction between technology and population accelerated the pace of technological progress permitting a take-off to the Post-Malthusian regime. The expansion of resources was partially counterbalanced by the enlargement of population and the economy was characterized by rapid growth rates of income per capita and population. The acceleration in technological progress increased the demand for human capital, while having two opposing effects on population growth. On the one hand, it eased households' budget constraints, allowing the allocation of more resources for raising children. On the other hand, it induced a reallocation of these additional resources toward child quality. In the Post-Malthusian regime, due to the limited demand for human capital, the first effect dominated and the rise in real income permitted households to increase their family size as well the quality of each child.

As investment in human capital took place the Malthusian steady state equilibrium vanished and the economy started to be attracted by the gravitational forces of the Modern Growth Regime. The interaction between investment in human capital and technological progress generated a virtuous circle: human capital generated faster technological progress, which in turn further raised the demand for human capital, inducing further investment in child quality, and ultimately initiating a demographic transition.[78] The offsetting effect of population growth on the growth rate of income per capita was eliminated and the interaction between human capital accumulation and technological progress permitted a transition to a state of sustained economic growth.

The theory suggests that the transition from stagnation to growth is an inevitable by-product of the Malthusian interaction between population and technology and its ultimate impact on the demand for human capital and the demographic transition . The timing of the transition differ however across countries and regions due to initial variations in geographical factors and historical accidents and their manifestation in variations in institutional, demographic, and cultural factors, trade patterns, colonial status, as well as in disparity in public policy.[79]

4.1.1 Central Building Blocks

The theory is based upon the interaction between the several building blocks: the Malthusian elements, the engines of technological progress, the origin of human capital formation, and the determinants of parental choice regarding the quantity and quality of offspring.

First, the Malthusian elements. Individuals are subjected to a subsistence consumption constraint and as long as the constraint is binding, an increase in income results in an increase in population growth. Technological progress, which brings about temporary gains in income per capita , triggers

[78] In less developed countries the stock of human capital determines the pace of adaptation of existing technologies whereas in developed countries it determined the pace of the advancement of the technological frontier.

[79] A fertile land in a Malthusian environment, for instance, would generate a larger population density and a scale effect.

therefore in early stages of development an increase in the size of the population that offsets the gain in income per capita due to the existence of decreasing returns to labor. Growth in income per capita is generated ultimately, despite decreasing returns to labor, since technological progress outpaces the rate of population growth.

Second, the forces behind technological progress in the process of development. The size of the population stimulates technological progress in early stages of development (Boserup (1965)), whereas investment in human capital is the prime engine of technological progress in more advanced stages of development. In the Malthusian era, the technological frontier was not distant from the working environment of most individuals, and the scale of the population affected the rate of technological progress due to its effect on: (a) the supply of innovative ideas, (b) the demand for new technologies, (c) technological diffusion, (d) the division of labor, and (e) trade.[80] As the distance from the technological frontier gets larger, however, the role of human capital becomes more significant in technological advancement (e.g., Nelson and Phelps (1966)) and individuals with high levels of human capital are more likely to advance the technological frontier.

Third, the origin of human capital formation. The introduction of new technologies is mostly skill-biased although in the long run, these technologies may be either "skill biased" or "skill saving." The "disequilibrium" brought about by technological change raises the demand for human capital.[81] Technological progress reduces the adaptability of existing human capital for the new technological environment and educated individuals have a comparative advantage in adapting to the new technological environment.[82]

Fourth, the determination of paternal decisions regarding the quantity and quality of their offspring. Individuals choose the number of children and their quality in the face of a constraint on the total amount of time that can be devoted to child-raising and labor market activities. The rise in the demand for human capital induces parents to substitute quality for quantity of children.[83]

4.1.2 The Basic Structure of the Model

Consider an overlapping-generations economy in which activity extends over infinite discrete time. In every period the economy produces a single homogeneous good using land and efficiency units of labor as inputs. The supply of land is exogenous and fixed over time whereas the number of efficiency units of labor is determined by households' decisions in the preceding period regarding the number and level of human capital of their children.

[80]The positive effect of the scale of the population on technological progress in the Malthusian epoch is supported by Boserup (1965) and recent evidence (e.g., Kremer (1993)). The role of the scale of the population in the modern era is, however, controversial. The distance to the technological frontier is significantly larger and population size per-se may have an ambiguous effect on technological progress, if it comes on the account of population quality.

[81]If the return to education rises with the *level* of technology the qualitative results would not be affected. Adopting this mechanism, however, would be equivalent to assuming that changes in technology were skill-biased throughout human history. Although on average technological change may have been skilled biased, Galor and Weil's mechanism is consistent with periods in which the characteristics of new technologies could be defined as unskilled-biased (most notably, the first phase of the industrial revolution).

[82]Schultz (1975) cites a wide range of evidence in support of this theory. Similarly, Foster and Rosenzweig (1996) find that technological change during the green revolution in India raised the return to schooling, and that school enrollment rates responded positively to this higher return. This element is central in the analysis of Galor and Tsiddon (1997), Galor and Moav (2000), and Hassler and Mora (2000).

[83]The existence of a trade-off between quantity and quality of children is supported empirically (e.g., Hanushek (1992) and Rosenzweig and Wolpin (1980)).

Production of Final Output Production occurs according to a constant-returns-to-scale technology that is subject to endogenous technological progress. The output produced at time t, Y_t, is

$$Y_t = H_t^{\alpha}(A_t X)^{1-\alpha}, \tag{1}$$

where H_t is the aggregate quantity of efficiency units of labor employed in period t, X is land employed in production in every period t, A_t represents the endogenously determined technological level in period t, $A_t X$ are therefore the "effective resources" employed in production in period t, and $\alpha \in (0,1)$.

Output per worker produced at time t, y_t, is

$$y_t = h_t^{\alpha} x_t^{1-\alpha}, \tag{2}$$

where $h_t \equiv H_t/L_t$ is the level of efficiency units of labor per worker, and $x_t \equiv (A_t X)/L_t$ is the level of effective resources per worker at time t.

Suppose that there are no property rights over land.[84] The return to land is therefore zero, and the wage per efficiency unit of labor is therefore equal to the output per efficiency unit of labor.

$$w_t = (x_t/h_t)^{1-\alpha}. \tag{3}$$

Preferences and Budget Constraints In each period t, a generation that consists of L_t identical individuals joins the labor force. Each individual has a single parent. Members of generation t (those who join the labor force in period t) live for two periods. In the first period of life (childhood), $t-1$, individuals consume a fraction of their parental unit time endowment. The required time increases with children's quality. In the second period of life (parenthood), t, individuals are endowed with one unit of time, which they allocate between child rearing and labor force participation. They choose the optimal mixture of quantity and quality of (surviving) children and supply their remaining time in the labor market, consuming their wages.

Individuals' preferences are represented by a utility function defined over consumption above a subsistence level $\tilde{c} > 0$, as well as over the quantity and quality (measured by human capital) of their (surviving) children.[85]

$$u^t = (c_t)^{1-\gamma}(n_t h_{t+1})^{\gamma} \qquad \gamma \in (0,1), \tag{4}$$

where c_t is the consumption of individual of generation t, n_t is the number of children of individual t, and h_{t+1} is the level of human capital of each child.[86] The utility function is strictly monotonically increasing and strictly quasi-concave, satisfying the conventional boundary conditions that assure that, for a sufficiently high income, there exists an interior solution for the utility maximization problem.

[84] The modeling of the production side is based upon two simplifying assumptions. First, capital is not an input in the production function, and second the return to land is zero. Alternatively it could have been assumed that the economy is small and open to a world capital market in which the interest rate is constant. In this case, the quantity of capital will be set to equalize its marginal product to the interest rate, while the price of land will follow a path such that the total return on land (rent plus net price appreciation) is also equal to the interest rate. Allowing for capital accumulation and property rights over land would complicate the model to the point of intractability, but would not affect the qualitative results.

[85] For simplicity parents derive utility from the expected number of surviving offspring and the parental cost child rearing is associated only with surviving children. A more realistic cost structure would not affect the qualitative features of the theory.

[86] Alternatively, the utility function could have been defined over consumption above subsistence rather than over a consumption set that is truncated from below by the subsistence consumption constraint. In particular, if $u^t = (c_t - \tilde{c})^{(1-\gamma)}(n_t h_{t+1})^{\gamma}$. the qualitative analysis would not be affected, but the complexity of the dynamical system would be greatly enhanced. The income expansion path would be smooth, transforming continuously from being nearly vertical for low levels of potential income to asymptotically horizontal for high levels of potential income. The subsistence consumption constraint would therefore generate the Malthusian effect of income on population growth at low income levels.

However, for a sufficiently low level of income the subsistence consumption constraint is binding and there is a corner solution with respect to the consumption level.[87]

Individuals choose the number of children and their quality in the face of a constraint on the total amount of time that can be devoted to child-raising and labor market activities. For simplicity, only time is required in order to produce child quantity and quality.[88] Let $\tau + e_{t+1}$ be the time cost for a member i of generation t of raising a child with a level of education (quality) e_{t+1}. That is, τ is the fraction of the individual's unit time endowment that is required in order to raise a child, regardless of quality, and e_{t+1} is the fraction of the individual's unit time endowment that is devoted for the education of each child.[89]

Consider members of generation t who are endowed with h_t efficiency units of labor at time t. Define potential income, z_t, as the potential earning if the entire time endowment is devoted to labor force participation, earning the competitive market wage, w_t, per efficiency unit. The potential income, $z_t \equiv w_t h_t$, is divided between consumption, c_t, and expenditure on child rearing (quantity as well as quality), evaluated according to the value of the time cost, $w_t h_t[\tau + e_{t+1}]$, per child. Hence, in the second period of life (parenthood), the individual faces the budget constraint

$$w_t h_t n_t(\tau + e_{t+1}) + c_t \leq w_t h_t \equiv z_t. \tag{5}$$

The Production of Human Capital Individuals' level of human capital is determined by their quality (education) as well as by the technological environment. Technological progress is assumed to raise the value of education in the production of human capital.[90] Technological progress reduces the adaptability of existing human capital for the new technological environment (the 'erosion effect'). Education, however, lessens the adverse effects of technological progress. That is, skilled individuals have a comparative advantage in adapting to the new technological environment. In particular, the time required for learning the new technology diminishes with the level of education and increases with the rate of technological change.

The level of human capital of children of a member i of generation t, h^i_{t+1}, is an increasing strictly concave function of their parental time investment in education, e^i_{t+1}, and a decreasing strictly convex function of the rate of technological progress, g_{t+1} :

$$h_{t+1} = h(e_{t+1}, g_{t+1}), \tag{6}$$

where $g_{t+1} \equiv (A_{t+1} - A_t)/A_t$. Education lessens the adverse effect of technological progress. That is, technology complements skills in the production of human capital (i.e., $h_{eg}(e^i_{t+1}, g_{t+1}) > 0$). In the absence of investment in quality, each individual has a basic level human capital that is normalized to 1 in a stationary technological environment, i.e., $h(0,0) = 1$.[91]

[87] The subsistence consumption constraint generates the positive income elasticity of population growth at low income levels, since higher income allows individuals to afford more children.

[88] If both time and goods are required in order to produce child quality, the process we describe would be intensified. As the economy develops and wages increase, the relative cost of a quality child will diminish and individuals will substitute quality for quantity of children.

[89] τ is assumed to be sufficiently small so as to assure that population can have a positive growth rate. That is, $\tau < \gamma$.

[90] Schultz [1975] cites a wide range of evidence in support of this assumption. More recently, Foster and Rosenzweig [1996] find that technological change during the green revolution in India raised the return to schooling, and that school enrollment rate responded positively to this higher return. The effect of technological transition on the return to human capital is at the center of the theoretical approach of Galor and Tsiddon (1997), Caselli (1999), Galor and Moav (2000), and Hassler and Rodriguez Mora (2002).

[91] For simplicity, investment in quality is not beneficial in a stationary technological environment, i.e., $h_e(0,0) = 0$, and in the absence of investment in education, there exists a sufficiently rapid technological progress, that due to the erosion effect renders the existing human capital obsolete (i.e., $\lim_{g\to\infty} h(0, g_{t+1}) = 0$). Furthermore, although the potential

Optimization Members of generation t choose the number and quality of their children, and therefore their own consumption, so as to maximize their intertemporal utility function subject to the subsistence consumption constraint. Substituting (5)-(6) into (4), the optimization problem of a member of generation t is:

$$\{n_t, e_{t+1}\} = argmax\{w_t h_t[1 - n_t(\tau + e_{t+1})]\}^{1-\gamma}\{(n_t h(e_{t+1}, g_{t+1})\}^{\gamma} \tag{7}$$

Subject to:

$$w_t h_t[1 - n_t(\tau + e_{t+1})] \geq \tilde{c};$$
$$(n_t, e_{t+1}) \geq 0.$$

Hence, as long as potential income at time t is sufficiently high so as to assure that $c_t > \tilde{c}$ (i.e., as long as $z_t \equiv w_t h_t$, is above the level of potential income at which the subsistence constraint is just binding, (i.e., $z_t > \tilde{z} \equiv \tilde{c}/(1-\gamma)$)), the fraction of time spent by individual t raising children is γ, while $1-\gamma$ is devoted for labor force participation. However, if $z_t \leq \tilde{z}$, the subsistence constraint is binding, the fraction of time necessary to assure subsistence consumption, $\tilde{c}$, is larger than $1-\gamma$ and the fraction of time devoted for child rearing is therefore below γ. That is,

$$n_t[\tau + e_{t+1}] = \begin{cases} \gamma & \text{if} \quad z_t \geq \tilde{z} \\ \\ 1 - [\tilde{c}/w_t h_t] & \text{if} \quad z_t \leq \tilde{z}. \end{cases} \tag{8}$$

Figure 4.1 shows the effect of an increase in potential income z_t on the individual's allocation of time between child rearing and consumption. The income expansion path is vertical as long as the subsistence consumption constraint is binding. As the wage per efficiency unit of labor increases in this income range, the individual can generate the subsistence consumption with a smaller labor force participation and the fraction of time devoted to child rearing increases. Once, the level of income is sufficiently high such that the subsistence consumption constraint is not binding, the income expansion path becomes horizontal at a level γ in terms of time devoted for child rearing.

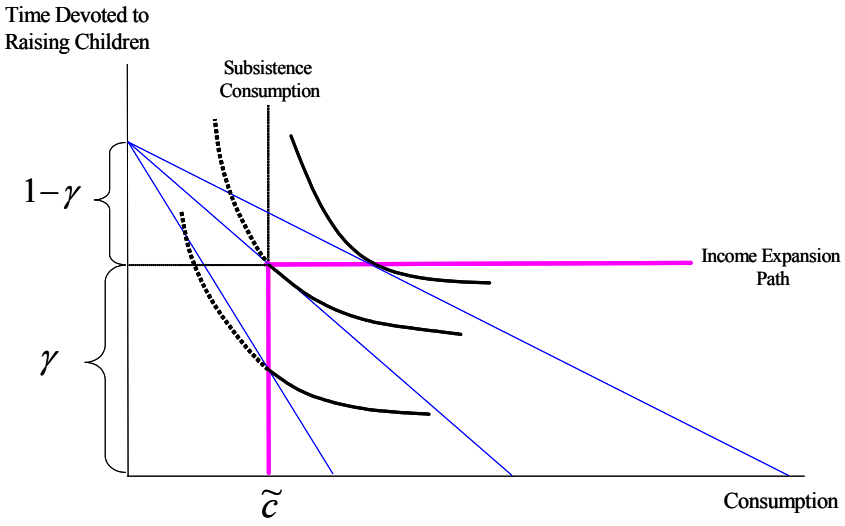


Figure 4.1. Preferences, Constraints, and Income Expansion Path

number of efficiency units of labor is diminished due to the transition from the existing technological state to a superior one (due to the erosion effect), each individual operates with a superior level of technology and the productivity effect is assumed to dominate. That is, $\partial y_t/\partial g_t > 0$.

Furthermore, the optimization with respect to e_{t+1} implies that the level of education chosen by members of generation t for their children, e_{t+1}, is an increasing function of g_{t+1}.

$$e_{t+1} = e(g_{t+1}) \quad \begin{cases} = 0 & \text{if} \quad g_{t+1} \leq \hat{g} \\ \\ > 0 & \text{if} \quad g_{t+1} > \hat{g} \end{cases} \tag{9}$$

where $e'(g_{t+1}) > 0$ and $e''(g_{t+1}) < 0 \ \ \forall g_{t+1} > \hat{g} > 0$.[92] Hence, regardless of whether potential income is above or below $\tilde{z}$, increases in wages will not change the division of child-rearing time between quality and quantity. However, the division between time spent on quality and time spent on quantity is affected by the rate of technological progress, which changes the return to education.

Furthermore, substituting (9) into (8), it follows that n_t is:

$$n_t = \begin{cases} \frac{\gamma}{\tau + e(g_{t+1})} \equiv n^b(g_{t+1}) & \text{if} \quad z_t \geq \tilde{z} \\ \\ \frac{1-[\tilde{c}/z_t]}{\tau + e(g_{t+1})} \equiv n^a(g_{t+1}, z(e_t, g_t, x_t)) & \text{if} \quad z_t \leq \tilde{z}. \end{cases} \tag{10}$$

where $z_t \equiv w_t h_t = z(e_t, g_t, x_t)$ as follows from (3) and (6)

Hence, as follows from the properties of $e(g_{t+1}), n^b(g_{t+1})$, and $n^a(g_{t+1}, z_t)$:
(a) An increase in the rate of technological progress reduces the number of children and increases their quality, i.e.,

$$\partial n_t / \partial g_{t+1} \leq 0 \text{ and } \partial e_{t+1} / \partial g_{t+1} \geq 0.$$

(b) If the subsistence consumption constraint is binding (i.e., if parental potential income is below $\tilde{z}$), an increase in parental potential income raises the number of children, but has no effect on their quality, i.e.,

$$\partial n_t / \partial z_t > 0 \text{ and } \partial e_{t+1} / \partial z_t = 0 \quad \text{if } z_t < \tilde{z}.$$

(c) If the subsistence consumption constraint is not binding (i.e., if parental potential income is above $\tilde{z}$), an increase in parental potential income does not affect the number of children or their quality, i.e.,

$$\partial n_t / \partial z_t = \partial e_{t+1} / \partial z_t = 0 \quad \text{if } z_t > \tilde{z}.$$

Technological Progress Suppose that technological progress, g_{t+1}, that takes place between periods t and $t+1$ depends upon the education per capita among the working generation in period t, e_t, and the population size in period t, L_t.[93]

$$g_{t+1} \equiv \frac{A_{t+1} - A_t}{A_t} = g(e_t, L_t), \tag{11}$$

where for $e_t \geq 0$ and a sufficiently large population size L_t, $g(0, L_t) > 0$, $g_i(e_t, L_t) > 0$, and $g_{ii}(e_t, L_t) < 0$, $i = e_t, L_t$.[94] Hence, for a sufficiently large population size, the rate of technological progress between

[92] $e''(g_{t+1})$ depends upon the third derivatives of the production function of human capital. $e''(g_{t+1})$ is assumed to be concave, which appears plausible.

[93] While the role of the scale effect in the Malthusian epoch, is essential, none of the existing results depend on the presence or the absence of the scale effect in the modern era. The functional form of technological progress given in (11) can capture both the presence and the absence of the scale effect in the modern era. In particular, the scale effect can be removed, once investment in education is positive, assuming for instance that $lim_{L \to \infty} g_L(e_t, L) = 0$ for $e_t > 0$.

[94] For a sufficiently small population the rate of technological progress is strictly positive only every several periods. Furthermore, the number of periods that pass between two episodes of technological improvement declines with the size of population. These assumptions assure that in early stages of development the economy is in a Malthusian steady-state with zero growth rate of output per capita, but ultimately the growth rates is positive and slow. If technological progress would occur in every time period at a pace that increases with the size of population, the growth rate of output per capita would always be positive, despite the adjustment in the size of population.

time t and $t+1$ is a positive, increasing, strictly concave function of the size and level of education of the working generation at time t. Furthermore, the rate of technological progress is positive even if labor quality is zero.

The state of technology at time $t+1$, A_{t+1}, is therefore

$$A_{t+1} = (1 + g_{t+1})A_t, \tag{12}$$

where the state of technology at time 0 is given at a level A_0.

Population The size of population at time $t+1$, L_{t+1}, is

$$L_{t+1} = n_t L_t, \tag{13}$$

where L_t is the size of population at time t and n_t is the number of children per person; L_0 is given. Hence, given (10), the evolution of population over time is

$$L_{t+1} = \begin{cases} n^b(g_{t+1})L_t & \text{if} \quad z_t \geq \tilde{z} \\ \\ n^a(g_{t+1}, z(e_t, g_t, x_t))L_t & \text{if} \quad z_t \leq \tilde{z}. \end{cases} \tag{14}$$

Effective Resources The evolution of effective resources per worker, $x_t \equiv (A_t X)/L_t$, is determined by the evolution of population and technology. The level of effective resources per worker in period $t+1$ is

$$x_{t+1} = \frac{1 + g_{t+1}}{n_t} x_t, \tag{15}$$

where $x_0 \equiv A_0 X/L_0$ is given. Furthermore, as follows from (10) and (11)

$$x_{t+1} = \begin{cases} \frac{[1+g(e_t,L_t)][\tau+e(g(e_t,L_t))]}{\gamma} x_t \equiv \phi^b(e_t, L_t)x_t & \text{if} \quad z_t \geq \tilde{z} \\ \\ \frac{[1+g(e_t,L_t)][\tau+e(g(e_t,L_t))]}{1-[\tilde{c}/z(e_t,g_t,x_t)]} x_t \equiv \phi^a(e_t, g_t, x_t, L_t)x_t & \text{if} \quad z_t \leq \tilde{z}, \end{cases} \tag{16}$$

where $\phi^b_e(e_t, L_t) > 0$, and $\phi^a_x(e_t, g_t, x_t, L_t) < 0 \ \ \forall e_t \geq 0$.

4.1.3 The Dynamical System

The development of the economy is fully determined by a sequence $\{e_t, g_t, x_t, L_t\}_{t=0}^{\infty}$ that satisfies (9), (11), 14), and (16), in every period t and describe the joint evolution of education, technological progress, effective resources per capita, and population over time.

The dynamical system is characterized by two regimes. In the first regime the subsistence consumption constraint is binding and the evolution of the economy is governed by a four dimensional non-linear first-order autonomous system:

$$\begin{cases} x_{t+1} = \phi^a(e_t, g_t, x_t; L_t)x_t \\ e_{t+1} = e(g(e_t; L_t)) \\ g_{t+1} = g(e_t, L_t) \\ L_{t+1} = n^a(g(e_t, L_t),\ z(e_t, g_t, x_t))L_t, \end{cases} \quad \text{for } z_t \leq \tilde{z} \tag{17}$$

where the initial conditions e_0, g_0, x_0 are historically given.

In the second regime the subsistence consumption constraint is not binding and the evolution of the economy is governed by a three dimensional system:

$$\begin{cases} x_{t+1} = \phi^b(e_t, x_t; L)x_t \\ e_{t+1} = e(g(e_t; L)) \qquad \text{for } z_t \geq \tilde{z} \\ L_{t+1} = n^b(g(e_t, L_t))L_t. \end{cases} \tag{18}$$

In both regimes, however, the analysis of the dynamical system is greatly simplified by the fact that the evolution of e_t and g_t is independent of whether the subsistence constraint is binding, and by the fact that, for a given population size L, the joint evolution of e_t and g_t is determined independently of the x_t. The education level of workers in period $t+1$ depends only on the level of technological progress expected between period t and period $t+1$, while technological progress between periods t and $t+1$ depends only on the level of education of workers in period t. Thus the dynamics of technology and education can be analyzed independently of the evolution resources per capita.

A. The Dynamics of Technology and Education The evolution of technology and education, for a given population size L, is characterized by the sequence $\{g_t, e_t; L\}_{t=0}^{\infty}$ that satisfies in every period t the equations $g_{t+1} = g(e_t; L)$, and $e_{t+1} = e(g_{t+1})$. Although this dynamical sub-system consists of two independent one dimensional, non-linear first-order difference equations, it is more revealing to analyze them jointly.

In light of the properties of the functions $e(g_{t+1})$ and $g(e_t; L)$ this dynamical sub-system is characterized by three qualitatively different configurations, which are depicted in Figures 4.2.A, 4.3.A and 4.4.A. The economy shifts endogenously from one configuration to another as population increases and the curve $g(e_t; L)$ shifts upward to account for the effect of an increase in population.

In Figure 4.2.A, for a range of small population size, the dynamical system is characterized by globally stable steady-state equilibria, $(\bar{e}(L), \bar{g}(L)) = (0, g^l(L))$, where $g^l(L)$ increases with the size of the population while the level of education remains unchanged. In Figure 4.3.A, for a range of moderate population size, the dynamical system is characterized by three steady state equilibria, two locally stable steady-state equilibria: $(\bar{e}(L), \bar{g}(L)) = (0, g^l(L))$ and $(\bar{e}(L), \bar{g}(L)) = (e^h(L), g^h(L))$, and an interior unstable steady-state $(\bar{e}(L), \bar{g}(L)) \equiv (e^u(L), g^u(L))$, where $(e^h(L), g^h(L))$ and $g^l(L)$ increase monotonically with the size of the population. Finally, in Figure 4.4.A, for a range of large population sizes, the dynamical system is characterized by globally stable steady-state equilibria, $(\bar{e}(L), \bar{g}(L)) = (e^h(L), g^h(L))$,where $e^h(L)$ and $g^h(L)$ increases monotonically with the size of the population.

B. Global Dynamics This section analyzes the evolution of the economy from the Malthusian Regime, through the Post-Malthusian Regime, to the demographic transition and Modern Growth. The global analysis is based a sequence of phase diagrams that describe the evolution of the system, within each regime, for a given population size, and the transition between these regimes as population increases in the process of development. Each of the phase diagrams is a two dimensional projection in the plain $(e_t, x_t; L)$,of the three dimensional system in the space $\{e_t, g_t, x_t; L\}$.

The phase diagrams, depicted in Figure 4.2.B, 4.4.B, and 4.5.B contain three elements: the Malthusian Frontier, which separates the regions in which the subsistence constraint is binding from those where it is not; the XX locus, which denotes the set of all triplets $(e_t, g_t, x_t; L)$ for which effective resources per worker are constant; and the EE locus, which denotes the set of all pairs $(e_t, g_t; L)$ for which the level of education per worker is constant.

The Malthusian Frontier

As was established in (17) and (18) the economy exits from the subsistence consumption regime when potential income, z_t, exceeds the critical level $\tilde{z}$. This switch of regime changes the dimensionality of the dynamical system from three to two.

Let the *Malthusian Frontier* be the set of all triplets of $(e_t, x_t, g_t; L)$ for which individuals' income equal $\tilde{z}$.[95] Using the definitions of z_t and $\tilde{z}$, it follows from (3) and (6) that the The Malthusian Frontier, $MM \equiv \{(e_t, x_t, g_t; L) : x_t^{1-\alpha} h(e_t, g_t)^{\alpha} = \tilde{c}/(1-\gamma)\}$.

Let the *Conditional Malthusian Frontier* be the set of all pairs $(e_t, x_t; L)$ for which, conditional on a given technological level g_t, individuals incomes equal $\tilde{z}$. Following the definitions of z_t and $\tilde{z}$, equations (3) and (6) imply that the Conditional Malthusian Frontier, $MM_{|g_t}$, is $MM_{|g_t} \equiv \{(e_t, x_t; L) : x_t^{(1-\alpha)} h(e_t, g_t)^{\alpha} = \tilde{c}/(1-\gamma) \mid g_t\}$, where x_t *is a decreasing strictly convex function of* e_t along the $MM_{|g_t}$locus.

Hence, the Conditional Malthusian Frontier, as depicted in Figures 4.2.B-4.4.B, is a strictly convex, downward sloping, curve in the (e_t, x_t) space. Furthermore, it intersects the x_t axis and approaches asymptotically the e_t axis as x_t approaches infinity. The frontier shifts upward as g_t increases in the process of development.

The XX Locus

Let XX be the locus of all triplets $(e_t, g_t, x_t; L)$ such that the effective resources per worker, x_t, is in a steady-state: $XX \equiv \{(e_t, x_t, g_t; L) : \ x_{t+1} = x_t\}$.

As follows from (15), along the XX locus the growth rates of population and technology are equal. Above the Malthusian frontier, the fraction of time devoted to child-rearing is independent of the level of effective resources per worker. In this case, the growth rate of population will just be a negative function of the growth rate of technology, since for higher technology growth, parents will spend more of their resources on child quality and thus less on child quantity. Thus there will be a particular level of technological progress which induces an equal rate population growth. Since the growth rate of technology is, in turn, a positive function of the level of education, this rate of technology growth will correspond to a particular level of education, denoted $\hat{e}$. Below the Malthusian Frontier, the growth rate of population depends on the level of effective resources per capita, x, as well as on the growth rate of technology. The lower is x, the smaller the fraction of the time endowment devoted to child-rearing, and so the lower is population growth. Thus, below the Malthusian frontier, a lower value of effective resources per capita would imply that lower values of technology growth (and thus education) would be consistent with population growth being equal to technology growth. Thus, as drawn in Figures 4.2.B, 4.3.B, and 4.4.B, lower values of x are consistent with lower values of e on the part of the XX locus that is below the Malthusian Frontier.

If the subsistence consumption constraint is not binding, it follows from (16) that for $z_t \geq \tilde{z}$, there exists a unique value $0 < \hat{e}(L) < e^h(L)$, such that $x_t \in XX$.[96]

$$x_{t+1} - x_t \begin{cases} > 0 & \textit{if} \quad e_t > \hat{e}(L) \\ = 0 & \textit{if} \quad e_t = \hat{e}(L) \\ < 0 & \textit{if} \quad e_t < \hat{e}(L) \end{cases} \tag{19}$$

Hence, the XX Locus, as depicted in Figures 4.2.B, 4.3.B, and 4.4.B is a vertical line above the Condi-

[95] Below the Malthusian Frontier, the effect of income on fertility will be positive, while above the frontier there will be no effect of income on fertility. Thus the Malthusian Frontier separates the Malthusian and Post-Malthusian regimes, on the one hand, from the Modern Growth regime, on the other, and crossing this frontier is associated with the demographic transition.

[96] In order to simplify the exposition without affecting the qualitative nature of the dynamical system, the parameters of the model are restricted so as to assure that the XX Locus is non-empty when $z_t \geq \tilde{z}$. That is, $\hat{g} < (\gamma/\tau) - 1 < g(e^h(L_0), L_0)$.

tional Malthusian Frontier at a level $\hat{e}(L)$.

If the subsistence constraint is binding, the evolution of x_t, is based upon the rate of technological change, g_t, the effective resources per-worker, x_t as well as the quality of the labor force, e_t. Let $XX_{|g_t}$ be the locus of all pairs $(e_t, x_t; L)$ such that $x_{t+1} = x_t$, for a given level of g_t. That is, $XX_{|g_t} \equiv \{(e_t, x_t; L) : x_{t+1} = x_t \mid g_t\}$ It follows from (16 that *for* $z_t \leq \tilde{z}$, *and for* $0 \leq e_t \leq \hat{e}(L)$, *there exists a single-valued function* $x_t = x(e_t)$ *such that* $(x(e_t), e_t) \in XX_{|g_t}$.

$$x_{t+1} - x_t \begin{cases} < 0 & \textit{if} \quad (e_t, x_t) > (e_t, x(e_t)) \quad \textit{for} \quad 0 \leq e_t \leq \hat{e}(L), \\ = 0 & \textit{if} \quad x_t = x(e_t) \quad \textit{for} \quad 0 \leq e_t \leq \hat{e}(L), \\ > 0 & \textit{if} \quad [(e_t, x_t) < (e_t, x(e_t)) \quad \textit{for} \quad 0 \leq e_t \leq \hat{e}(L),] \quad \textit{or} \quad [e_t > \hat{e}(L)] \end{cases} \tag{20}$$

Hence, without loss of generality, the locus $XX_{|g_t}$ is depicted in Figure 4.2, as an upward slopping curve in the space (e_t, x_t), defined for $e_t \leq \hat{e}(L)$. $XX_{|g_t}$ is strictly below the Conditional Malthusian Frontier for value of $e_t < \hat{e}(L)$, and the two coincides at $\hat{e}(L)$. Moreover, the Conditional Malthusian Frontier, the XX Locus, and the $XX_{|g_t}$ Locus, coincide at $(\hat{e}(L), \hat{x}(L))$.

The EE Locus

Let EE be the locus of all triplets $(e_t, g_t, x_t; L)$ such that the quality of labor, e_t, is in a steady-state: $EE \equiv \{(e_t, x_t, g_t; L) : e_{t+1} = e_t\}$.

As follows from (9) and (11), $e_{t+1} = e(g(e_t; L)$ and thus, for a given population size, the steady-state values of e_t are independent of the values of x_t and g_t. The locus EE evolves through three phases in the process of development, corresponding to the three phases that describe the evolution of education and technology, as depicted in Figures 4.2.A, 4.3.A, and 4.4.A.

In early stages of development, when population size is sufficiently small, the joint evolution of education and technology is characterized by a globally stable temporary steady-state equilibrium, $(\bar{e}(L), \bar{g}(L)) = (0, g^l(L))$, as depicted in Figure 4.2.A. The corresponding EE Locus, depicted in the space $(e_t, x_t; L)$ in Figure 4.2.B, is vertical at the level $e = 0$, for a range of small population sizes. Furthermore, for this range, the global dynamics of e_t are given by:

$$e_{t+1} - e_t \begin{cases} = 0 & \textit{if} \quad e_t = 0 \\ < 0 & \textit{if} \quad e_t > 0. \end{cases} \tag{21}$$

In later stages of development as population size increases sufficiently, the joint evolution of education and technology is characterized by multiple locally stable temporary steady-state equilibria, as depicted in Figure 4.3.A. The corresponding EE Locus, depicted in the space $(e_t, x_t; L)$ in Figure 4.3.B, consists of 3 vertical lines corresponding the three steady-state equilibria for the value of e_t. That is, $e = 0$, $e = e^u(L)$, and $e = e^h(L)$. The vertical lines $e = e^u(L)$, and $e = e^h(L)$ shift rightward as population size increases. Furthermore, the global dynamics of e_t in this configuration are given by:

$$e_{t+1} - e_t \begin{cases} < 0 & \textit{if} \quad 0 < e_t < e^u(L) \quad \textit{or} \quad e_t > e^h(L) \\ = 0 & \textit{if} \quad e_t = (0,\ e^u(L),\ e^h(L)) \\ > 0 & \textit{if} \quad e^u(L) < e_t < e^h(L). \end{cases} \tag{22}$$

In mature stages of development when population size is sufficiently large, the joint evolution of education and technology is characterized by a globally stable steady-state equilibrium, $(\bar{e}(L), \bar{g}(L)) = (e^h(L), g^h(L))$, as depicted in Figure 4.4.A. The corresponding EE Locus, as depicted in Figure 4.4.B in the space $(e_t, x_t; L)$, is vertical at the level $e = e^h(L)$. This vertical line shifts rightward as population size increases. Furthermore, the global dynamics of e_t in this configuration are given by:

$$e_{t+1} - e_t \begin{cases} > 0 & if \quad 0 \leq e_t < e^h(L) \\ = 0 & if \quad e_t = e^h(L). \\ < 0 & if \quad e_t > e^h(L). \end{cases} \tag{23}$$

Conditional Steady-State Equilibria

In early stages of development, when population size is sufficiently small, the dynamical system, as depicted in Figure 4.2.B is characterized by a unique and globally stable conditional steady-state equilibrium.[97] It is given by a point of intersection between the EE Locus and the $x_{t+1} = x_t$ Locus. That is, conditional on a given technological level, g_t, the Malthusian steady-state $(0, \bar{x}(L))$ is globally stable.[98] In later stages of development as population size increases sufficiently, the dynamical system as depicted in Figure 4.3.B is characterized by two conditional steady-state equilibria. The Malthusian conditional steady-state equilibrium is locally stable, whereas the steady-state equilibrium $(e^u(L), x^u(L))$ is a saddle point.[99] For education levels above $e^u(L)$ the system converges to a stationary level of education $e^h(L)$ and possibly to a steady-state *growth rate* of x_t. In mature stages of development when population size is sufficiently large, the system convergence globally to an educational level $e^h(L)$ and possibly to a steady-state *growth rate* of x_t.

4.1.4 From Malthusian Stagnation to Sustained Growth

The economy evolves from an epoch of Malthusian stagnation through the Post-Malthusian regime to the demographic transition and a Modern Growth regime. This pattern and the prime driving forces in this transition emerge from the phase diagrams depicted in Figures 4.2-4.4.

Consider an economy in early stages of development. Population size is relatively small and the implied slow rate of technological progress does not provide an incentive to invest in the education of children. As depicted in Figure 4.2.A,the interaction between education, e_t, and the rate of technological change, g_t, for a constant small population, L^l, is characterized by a globally stable steady-state equilibrium $(0, g^l(L))$, where education is zero and the rate of technological progress is slow. This steady-state equilibrium corresponds to a globally stable conditional Malthusian steady-state equilibrium, depicted in Figure 4.2.B. For a constant small population, L^l,and for a given rate of technological progress, effective resources per capita, as well as the level of education are constant, and output per capita is therefore constant as well. Moreover, shocks to population or resources will be resolved in a classic Malthusian fashion.

As population grows slowly in reaction to technological progress, the $g(e_{t+1}, L^l)$ locus, depicted in Figure 4.2.A, gradually shifts upward and the steady-state equilibrium shifts vertically upward reflecting small increments in the rate of technological progress, while the level of education remains constant at a zero level. Similarly, the conditional Malthusian steady-state equilibrium drawn in Figure 4.2.B shifts vertically upward , as the XX locus shifts upward. However, output per capita remains initially constant at the subsistence level and ultimately creeps forward at a miniscule rate.

[97] Since the dynamical system is discrete, the trajectories implied by the phase diagrams do not necessarily approximate the actual dynamic path, unless the state variables evolve monotonically over time. As shown, the evolution of e_t is monotonic, whereas the evolution and convergence of x_t may be oscillatory. Non-monotonicity in the evolution of x_t may arise only if $e < \hat{e}$ and it does not affect the qualitative description of the system. Furthermore, if $\phi_x^a(e_t, g_t, x_t)x_t > -1$ the conditional dynamical system is locally non-oscillatory. The phase diagrams in Figures 4.3.A-4.5.A are drawn under the assumptions that assure that there are no oscillations.

[98] The local stability of the steady-state equilibrium $(0, \overline{x}(g_t))$ can be derived formally. The eigenvalues of the Jacobian matrix of the conditional dynamical system evaluated at the conditional steady-state equilibrium are both smaller than one (in absolute value)

[99] Convergence to the saddle point takes place only if the level of education is e^u. That is, the saddle path is the entire vertical line that corresponds to $e_t = e^u$.

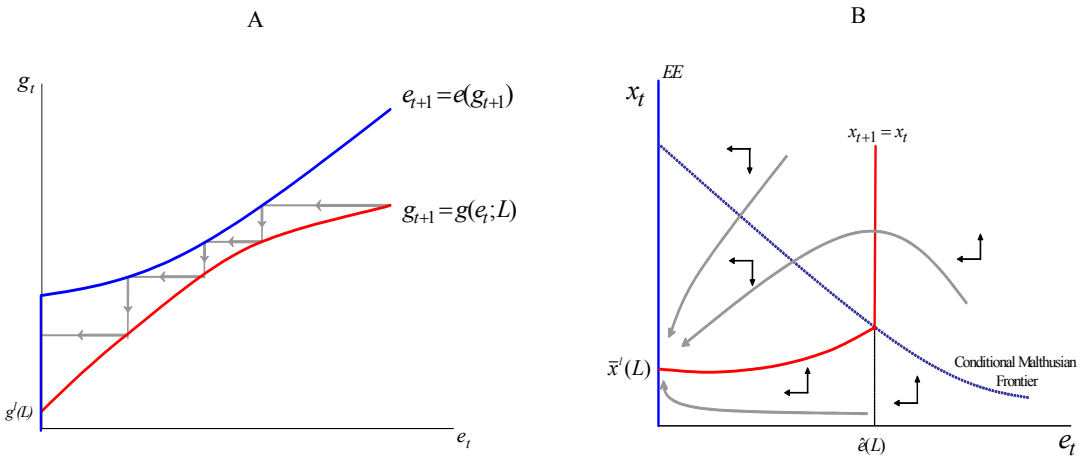


Figure 4.2. The Evolution of Technology, g_t, Education, e_t, and Effective Resources, x_t
Small Population

Over time, the slow growth in population that takes place in the Malthusian regime raises the rate of technological progress and shift the $g(e_{t+1}, L^l)$ locus in Figure 4.2.A sufficiently upward, generating a qualitative change in the dynamical systems depicted in Figure 4.3.A.

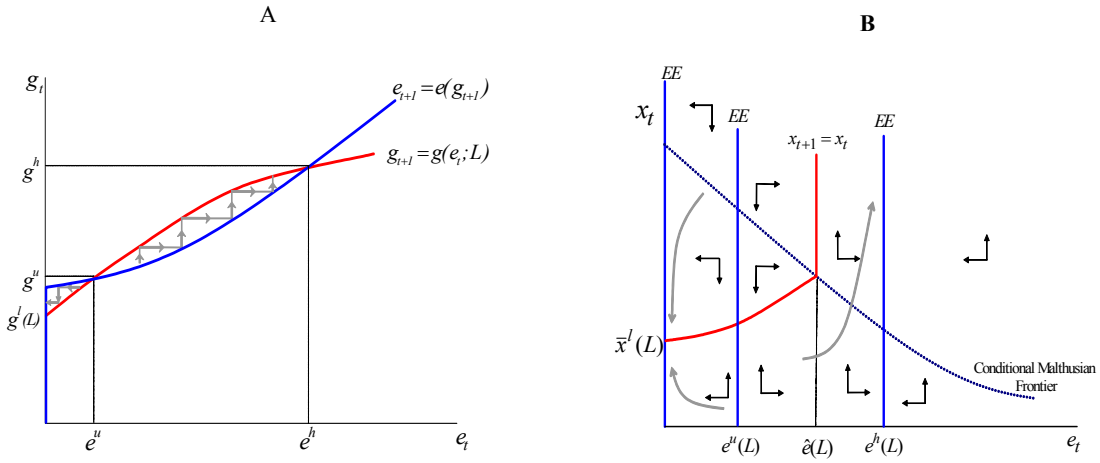


Figure 4.3. The Evolution of Technology, g_t, Education, e_t, and Effective Resources, x_t :
Moderate Population

The dynamical system of education and technology, for a moderate population, is characterized by multiple, history-dependent, stable steady state equilibria: The steady-state equilibria $(0, g^l(L))$ and $(e^h(L), g^h(L))$ are locally stable, whereas $(e^u(L), g^u(L))$ is unstable. Given the initial conditions, in the absence of large shocks the economy remains in the vicinity of the low steady-state equilibrium $(0, g^l(L))$, where education is still zero but the rate of technological progress is moderate. This steady-state equilibria correspond to a multiple locally stable conditional Malthusian steady-state equilibrium, depicted in Figure 4.3.B. A Malthusian steady-state, characterized by constant resources per capita, slow

technological progress, and no education, and a modern growth steady state, characterized by a high level of education, rapid technological progress, growing income per capita, and moderate population growth. However, since the economy starts in the vicinity of Malthusian steady state, it remains there.[100]

As the rate of technological progress continue to rise in reaction to the increasing population size, the $g(e_{t+1}, L_t)$ locus shifts upward further and ultimately, as depicted in Figure 4.4, the dynamical system experiences another qualitative change. The Malthusian steady state equilibrium vanishes, and the economy is the system is characterized by a unique globally stable modern steady-state equilibrium $(e^h(L), g^h(L))$ characterized by high levels of education and technological progress.

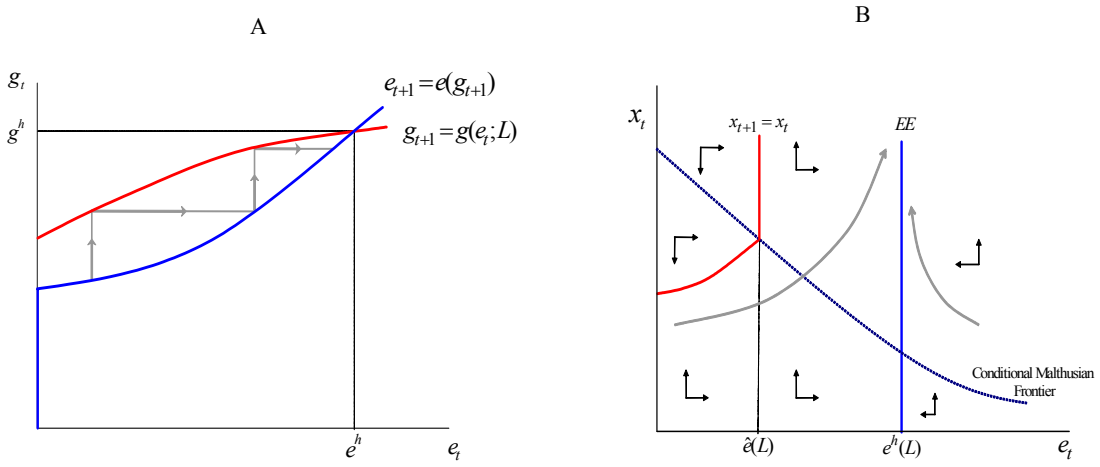


Figure 4.4. The Evolution of Technology, g_t, Education, e_t, and Effective Resources, x_t
Large Population

Increases in the rate of technological progress and the level of education feed back on each other until the economy converges rapidly to the stable modern steady state equilibrium. The increase in the pace of technological progress has two opposing effects on the evolution of population. On the one hand, it eased households' budget constraints, allowing the allocation of more resources for raising children. On the other hand, it induced a reallocation of these additional resources toward child quality. In the Post-Malthusian regime, due to the limited demand for human capital, the first effect dominated and the rise in real income permitted households to increase their family size as well the quality of each child.[101] The interaction between investment in human capital and technological progress generate a virtuous circle: human capital formation prompted faster technological progress, which in turn further raised the demand for human capital, inducing further investment in child quality, and ultimately, as the economy crosses the Malthusian frontier triggering a demographic transition. The offsetting effect of population growth on the growth rate of income per capita is eliminated and the interaction between human capital accumulation and technological progress permitted a transition to a state of sustained economic growth.

[100]Large shock to education or technological progress would permit the economy to jump to the Modern Growth steady state, but this possibility appears inconsistent with the evidence.

[101]Literally, income per capita does not change during the Post-Malthusian regime. It remains fixed at the subsistence level. This is an artifact of the assumption that the only input into child (quality and quantity) is parental time, and that this time input does not produce measured output. If child-rearing, especially the production of quality, requires goods or time supplied through a market (e.g., schooling), the shift toward higher child quality that takes place during the post-Malthusian regime would be reflected in higher market expenditures (as opposed to parental time expenditures) and rising measured income.

In the modern growth regime, resources per capita rise, as technological progress outstrips population growth. Provided that population size is constant (i.e., population growth is zero), the levels of education and technological progress and the growth rates of resources per capita, and thus the output per capita are constant in the modern growth steady state equilibrium.[102]

4.1.5 Major Hypotheses and their Empirical Validity

The theory generates several hypotheses about the evolution of population, human capital and income per capita in the process of development, underlying the roles of the inherent interaction between population and technology in the Malthusian epoch, as well as the formation of human capital in the second phase of the Industrial Revolution and the associated demographic transition, in the emergence of a state of sustained economic growth.

Main Hypotheses:

- During the initial phases of the Malthusian epoch the growth rate of output per capita is nearly zero and the growth rate of population is miniscule, reflecting the sluggish pace of technological progress and the full adjustment of population to the expansion of resources. In the later phases of the Malthusian epoch, the increasing rate of technological progress, along with the inherent delay in the adjustment of population to the rise in income per capita, generated a positive but very small growth rates of output per capita and population.

The hypothesis is consistent with the evidence provided in section 2.1 about the evolution of the world economy in the Malthusian epoch. In particular, the infinitesimal pace of resource expansion in the first millennium was reflected in a miniscule increase in the Western European population from 24.7 million people in the year 1 to 25.4 million in the year 1000, along with a zero average growth rate of output per capita. The more rapid (but still very slow) expansion of resources in the period 1000-1500, permitted the Western European population to grow at a slow average rate of 0.16% per year, from 25 million in the year 1000 to 57 million in the year 1500, along with a slow average growth rate of income per capita at a rate of about 0.13% per year. Resource expansion over the period 1500-1820 had a more significant impact on the Western European population that grew at an average pace of 0.26% per year, from 57 million in the year 1500 to 133 million in the year 1820, along with a slightly faster average growth rate of income per capita at a rate of about 0.15% per year.

- The reinforcing interaction between population and technology during the Malthusian epoch, increased the size of the population sufficiently so as to support a faster pace of technological progress, generating the transition to the Post-Malthusian Regime. The growth rates of output per capita increased significantly, but the positive Malthusian effect of income per capita on population growth was still maintained, generating a sizeable increase in population growth, and offsetting some of the potential gains in income per capita. Moreover, human capital accumulation did not play a significant role in the transition to the Post-Malthusian Regime and thus in the early take-off in the first phase of the Industrial Revolution. It emerged in the midst of the Post-Malthusian Regime, inducing further technological progress.

[102] If population growth is positive in the Modern Growth regime, then education and technological progress continue to rise, and, similarly, if population growth is negative they fall. In fact, the model makes no firm prediction about what the growth rate of population will be in the Modern Growth regime, other than that population growth will fall once the economy exits from the Malthusian region. If the growth rate of technology is related to the growth rate of population, rather than to its level, then there exists a steady state modern growth regime in which the growth rates of population and technology would be constant. Further, such a steady state would be stable: if population growth fell, the rate of technological progress would also fall, inducing a rise in fertility.

The hypothesis is consistent with the evidence provided in section 2.2 about the evolution of the world economy in the Post-Malthusian regime. In particular, the acceleration in the pace of resource expansion in the period 1820-1870 increased the Western European population from 133 million people in the year 1820 to 188 million in the year 1870, and the average growth rate of output per capita over this period increased significantly to 0.95% per year. Furthermore, the evidence suggest that the industrial demand for human capital increased only in the second phase of the of the Industrial Revolution. As shown by Clark (2003) human capital formation prior to the Industrial Revolution as well as in its first phase occurred in an era in which the market rewards for skill acquisition were at historically low levels.[103]

- The acceleration in the rate of technological progress increased the demand for human capital in the Post-Malthusian Regime, inducing significant investment in human capital, and triggering the demographic transition and a rapid pace of economic growth.

The hypotheses is consistent with the evidence, provided in section 2.3 and depicted partly in Figure 4.5, about the significant rise in the industrial demand for human capital in the second phase of the Industrial Revolution and a marked increased in educational attainment, in association with a decline in fertility rates, and a transition to a state of sustained economic growth. In particular, it is consistent with the revisionist view on the British Industrial Revolution (e.g., Crafts and Harley 1992, Clark 2001, and Voth 2003) that argue that the first phase of the Industrial Revolution in England was characterized by a moderate increase in the growth rate of output per capita, and the standard of living and the "take-off", as depicted in Figure 4.5, occurred only in 1860s.

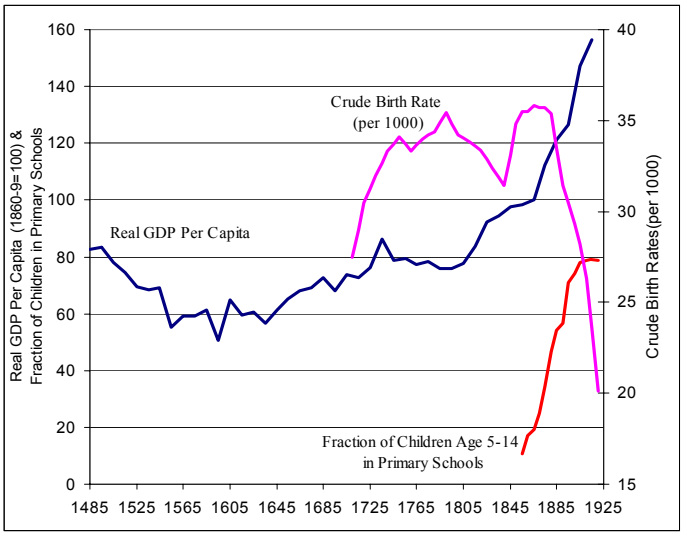


Figure 4.5. The Sharp Rise in Real GDP Per Capita and it association with Investment in Education and Fertility Decline: England 1865-1920
Source: Clark (2001), Feinstein (1972), Flora et al. (1983), Wrigley and Schofield (1981)

Furthermore, quantitative analysis of unified growth theories by Fernandez-Vilaverde (2003), Doepke (2004), and Pereira (2004) suggests that indeed the rise in the demand for human capital was a significant force behind the demographic transition and the emergence of a state of sustained economic growth[104]

[103]The rise in human capital formation over this period may reflect religious, cultural and social forces, as well as the rise in valuation for offspring quality due to the forces of natural selection, as discussed in section 5.2.

[104]The rise in the demand for human capital in Fernandez-Vilaverde (2003) is based on capital-skill complementarity, and is indistinguishable from the complementarity between technology and skills (in the short run) that is maintained by Galor and Weil (2000).

Moreover, the theory is consistent with the observed simultaneous onset of the demographic transition across Western European countries that differed significantly in their income per capita. It suggests that a universal rise in the demand for human capital in Western Europe (as documented in section 2.3.3) generated this simultaneous transition. It should be noted that the lack of clear evidence about the increase in the return to human capital in the second phase of the Industrial Revolution does not indicate the absence of a significant increase in the demand for human capital over this period. The significant increase in schooling that took place in the 19th century and in particular the introduction of public education (e.g., The Education Act of 1870) that lowered the cost of education, generated significant increase in the supply of educated workers that may have prevented a significant rise in the return to education.[105]

- a. The growth process is characterized by stages of development and the evolution of the growth rates of output per capita is nonlinear. Technological leaders experienced a monotonic increase in the growth rates of their income per capita. Their growth was rather slow in early stages of development, it increased rapidly during the take-off from the Malthusian epoch, and then continued to rise at a lower pace, possibly stabilizing at higher level. Technological followers that made the transition to sustained economic growth, in contrast, experienced a non-monotonic increase in the growth rates of their income per capita. Their growth rates rather slow in early stages of development, it increased rapidly in the early stages of the take-off from the Malthusian epoch and was boosted by the adoption of technologies from the existing technological frontier, and then once the economy reached the technological frontier, the growth rates dropped to the level of the technological leaders.

  b. The differential timing of the take-off between economies generated convergence clubs characterized by a group of poor countries in the vicinity of the Malthusian equilibrium, a group of rich countries in the vicinity of the sustained growth equilibrium, and others that are in the attempting to shift from one camp to another.[106]

These hypotheses are consistent the Maddison (2001)'s evidence about the growth process in the last 2000 years, as well as with contemporary cross section evidence that suggest that the growth process is characterized by multiple growth regimes (e.g., Duraluf and Johnson 1995) and thus with non-linearities in the growth process (e.g., Duraluf and Quah 1998, and Bloom, Canning and Silva 2003), and that divergence along with a twin peak have emerged in the distribution of income across countries in the world (Quah 1996, 1997, Jones 1997, and Prichett 1997).[107]

4.2 Complementary Theories

Subsequent theories of economic growth in the very long run demonstrate that the unified theory of economic growth can be augmented and fortified by additional characteristics of the transition from stagnation to growth without altering the fundamental hypothesis regarding the central roles played by the emergence of human capital formation and the demographic transition in this process. Various qualitative and quantitative unified theories explore plausible mechanisms for the emergence of human capital and the onset of the demographic transition such as, the rise in the demand for human capital

[105]Some of this supply response was a direct reaction of the potential increase in the return to human capital, and thus may only operate to partially offset the increase in the return to human capital, but the reduction in the cost of education via public schooling, generated an additional force that operated towards a reduction in the return to human capital.

[106]For the definition and the theory of club convergence see Azariadis (1996) and Galor (1996).

[107]Other studies that focused on nonlinearity of the growth process includes Fiaschi and Lavezzi (2003). Other research on the emergence of twin peak is Feyrer (2003).

(due to: technological acceleration, capital-skill complementarity, skilled biased technological change, and reallocation of resources towards skilled intensive sectors), the decline in child and infant mortality, the rise in life expectancy, the emergence of public education, the decline in child labor, as well as cultural and genetic evolution in the valuation of human capital. They suggest that indeed the emergence of human capital formation, and the onset of the demographic transition played a central role in the shift from stagnation to growth.

4.2.1 Alternative Mechanisms for the Emergence of Human Capital Formation

The emergence of human capital formation and its impact on the demographic transition and the technological frontier is a central element in the transition from the Post-Malthusian Regime to the state of sustained economic growth in all unified theories of economic growth in which population, technology and income per capita are endogenously determined.[108] Various complementary mechanisms that generate or reinforce the rise in human capital formation have been proposed and examined quantitatively, demonstrating the robustness and the empirical plausibility of this central hypothesis.

The Rise in the Industrial Demand for Human Capital

The rise in industrial demand for human capital in advanced stages of industrialization, as documented in section 2.3.3, and its impact on human capital formation led researchers to incorporate it as a central feature in unified theories of economic growth.

The link between industrial development and the demand for human capital have been modeled in various complementary ways. Galor and Weil (2000) modeled the rise in the demand for human capital as an outcome of the acceleration in technological progress, underlying the role of educated individuals in coping with a rapidly changing technological environment. Their mechanism is founded on the premise that the introduction of new technologies increases the demand for skilled labor in the short-run, although in some periods the characteristics of new technologies may be complementarity to unskilled labor, as was the case in the first phase of the Industrial Revolution.[109]

Subsequent unified theories of economic growth have demonstrated that the rise in the demand for human capital in association with advanced stages of industrialization could emerge from alternative mechanisms, without altering the fundamental insights of the theory. Fernandez-Vilaverde (2003) bases his quantitative unified theory on capital-skill complementarity, Doepke (2004) constructs his unified theory on the basis of a rising *level* of skilled-intensive industrial technology, and Galor and Mountford (2003) generate the rise in the demand for human capital via an increased specialization in the production of skilled-intensive goods due to international trade.

The rise in the demand for human capital stimulated public policy designed to enhance investment in human capital. In particular, as established in the quantitative unified theory of Doepke (2004), educational policy and child labor laws in England played an important role in human capital formation and the demographic transition.

Mortality Decline, the Rise in Life Expectancy, and Human Capital Formation

Several unified theories of economic growth demonstrated that the basic mechanism for the emergence of human capital proposed by Galor and Weil (2000) can be augmented and reinforced by the incorporation of the effect of the decline in mortality rates and the rise in life expectancy (as

[108] Even in the multiple-regime structure of Lucas (2002) a shock to the return to human capital is suggested in order to generate the switch from the Malthusian Regime to the Modern Growth Regime.

[109] Evidence for the complementarity between technological progress (or capital) and skills is provided by Katz and Goldin (1988) and on the basis of cross-section studies by Duffy, Papageorgiou and Perez-Sebastain (2004).

documented in section 2.3.2) on the rise in human capital formation, the decline in the desirable number of surviving offspring, and thus on the transition from stagnation to growth.[110]

The significant decline in mortality rates in developed countries since the 18th century, as depicted in Figure 2.24, and the recent decline in mortality rates in less developed countries, as depicted in Figure 2.25, corresponded to an acceleration in the rise in life expectancy and a significant rise in human capital formation, towards the end of the 19th century in developed countries (Figures 2.26 and 2.28) and towards the middle of the 20th century in less developed countries (Figures 2.27 and 2.31). The rise in the expected length of the productive life may have increased the potential rate of return to investments in children's human capital, and thus could have induced an increase in human capital formation along with a decline in fertility. However, despite the gradual rise in life expectancy in developed and less developed countries, investment in human capital has been insignificant as long as the industrial demand for human capital has not emerged. Thus, it appears that the industrial demand for human capital, as documented in section 2.3.3, provided the inducement for investment in education and the associated reduction in fertility rates, whereas the prolongation of life may have re-enforced and complemented this process.

Galor and Weil (1999) argue that the Malthusian interaction between technology and population accelerated the pace of technological progress, improving industrial technology as well as medical and health technologies. Consistent with the historical evidence provided in section 2.3.3, the improvements in the industrial technology increased the demand for human capital, whereas the development of medical technology and health infrastructure generated a significant rise in life expectancy. The expected rate of return to human capital investment increased therefore due to the prolongation of life, as well as the rise in industrial demand for human capital, enhancing the positive interaction between schooling and technological progress, bringing about a demographic transition and the emergence of the state of sustained economic growth.

Various theories formally examined mechanisms that capture the interaction between human capital formation, the decline in mortality rate, and the rise in life expectancy, in the process of development.[111] Cervellati and Sunde (2003) and Boucekkine, de la Croix and Licandro (2003) focus on the plausible role of the reinforcing interaction between life expectancy and human capital formation in the transition from stagnation to growth, abstracting from its effect on fertility decisions. Others suggest that a decline in mortality rates increased the return to investment in human capital via: (a) increased population density and thus efficiency of the transmission of human capital (Lagerlof 2003a), (b) increased population growth and the advancement of skill-biased technologies (Weisdorf 2004), and (c) improved healthiness and thus the capacity to absorb human capital (Hazan and Zoabi 2004), generating a substitution of quality for quantity, a demographic transition and a transition to a state of sustained economic growth.[112]

Capital-Skill Complementarity and the Emerging Incentives for Capitalists to Support Education Reforms

The accumulation of physical capital in the early stages of industrialization enhanced the importance of human capital in the production process and generated an incentive for the capitalists to support the provision of public education for the masses.[113] Consistent with the evidence provided in

[110]The effect of an increase in life expectancy on the incentive of individuals to invest in their own human capital is well established since Ben-Porath (1967). See Kalemli-Ozcan, Ryder and Weil (2000), as well.

[111]As argued in section 3.2.3, qualitative and quantitative evidence do not lend credence to the theory that a decline in infant and child mortality rates *triggered* the decline in the number of surviving offspring and the increase in the investment in offspring's human capital.

[112]See Iyigun (2003) as well.

[113]Attentively, other argued that increased polarization induced the elite to enact costly educational reforms. Grossman

section 2.3.3, Galor and Moav (2004) argue that due to capital-skill complementarity, the accumulation of physical capital by the capitalists in the first phase of the Industrial Revolution increased the importance of human capital in sustaining the rate of return to physical capital inducing capitalists to support the provision of public education for the masses.[114]

The Decline in Child Labor

Other theories that focused on the transition from stagnation to growth suggested that the central role of human capital formation and the demographic transition can be augmented and reinforced by the incorporation of the adverse effect of the rise in the demand for human capital on child labor. Hazan and Berdugo (2002) suggest that technological change increased the wage differential between parental labor and child labor inducing parents to reduce the number of their children and to further invest in their quality, stimulating human capital formation, a demographic transition, and a shift to a state of sustained economic growth.[115] Alternatively, the rise in the importance of human capital in the production process, as documented in section 2.3.3, induced industrialists to support laws that abolish child labor (Doepke and Zilibotti (2003)), inducing a reduction in child labor, and stimulating human capital formation and a demographic transition.

Cultural and Genetic Evolution in the Valuation of Human Capital

Human capital formation and its impact on the decline in the desirable number of surviving offspring may have been reinforced by cultural or genetic evolution in the attitude of individuals towards human capital formation. Consistent with the gradual rise in literacy rates prior to the Industrial Revolution, Galor and Moav (2002), argue that during the epoch of Malthusian stagnation that had characterized most of human existence, individuals with a higher valuation for offspring quality generated an evolutionary advantage and their representation in the population gradually increased. The increase in the rate of return to human capital along with the increase in the bias towards quality in the population reinforced the substitution towards child quality, setting the stage for a significant increase in human capital formation along with a rapid decline in fertility.

4.2.2 Alternative Triggers for the Demographic Transition

The demographic transition that separated the Post-Malthusian Regime and the Sustained Growth Regime is a central element in quantitative and qualitative unified theories of economic growth in which population, technology and income per capita are endogenously determined. As discussed in section 2.3.2, the demographic transition brought about a reversal in the unprecedented increase in population growth that occurred during the Post-Malthusian Regime, leading to a significant reduction in fertility rates and population growth in various regions of the world, and enabling economies to convert a larger share of the fruits of factor accumulation and technological progress into growth of output per capita.[116] The demographic transition enhanced the growth process reducing the dilution of the stock of capital and land, enhancing the investment in the human capital of the population, and alternating the age

and Kim (1999) argue that education decreases predation, and Bowles and Gintis (1975) suggest that educational reforms are designed to *sustain* the existing social order, by displacing social problems into the school system. In contrast, Bourguignon and Verdier (2000) suggest that if political participation is determined by the education (socioeconomic status) of citizens, the elite may not find it beneficial to subsidize universal public education despite the existence of positive externalities from human capital.

[114]Since firms have limited incentive to invest in the general human capital of their workers, in the presence of credit market imperfections, the level of education would be suboptimal unless it would be financed publicly (Galor and Zeira (1993), Duraluf (1996), Fernandez and Rogerson (1996) Benabou (2000), Mookherjee and Ray (2003), and Galor and Moav (2004)). Moreover, a mixture of vocational and general education would be enacted (Bertocchi and Spagat 2004).

[115]The decline in the relative wages of children is documented empirically (e.g., Horrell and Humphries (1995)).

[116]Demographic shocks generate a significant effect on economic growth in Connolly and Peretto (2003) as well.

distribution of the population, increasing temporarily the size of the labor force relative to the population as a whole.[117]

Various complementary mechanisms for the demographic transition have been proposed in the context of unified growth theories, establishing, theoretically and quantitatively the importance of this central hypothesis in the understanding of the transition from stagnation to growth.[118]

The Emergence of Human Capital Formation

The gradual rise in the demand for human capital in the process of industrialization, as documented in section 2.3.3, and its close association with the timing of the demographic transition has led researchers to argue that the increasing role of human capital in the production process induced households to increase their investment in the human capital of their offspring, ultimately leading to the onset of the demographic transition.

The link between the rise in the demand for human capital and the demographic transition have been modeled in various complementaries ways. Galor and Weil (2000) argue that the gradual rise in the demand for human capital induced parents to invest in the human capital of their offspring. In the early stages of the transition from the Malthusian regime, the effect of technological progress on parental income permitted the rise in population growth as well as the average quality. Further increases in the rate of technological progress ultimately induced a reduction in fertility rates, generating a demographic transition in which the rate of population growth declined along with an increase in the average level of education. Thus, consistent with historical evidence, the theory suggests that prior to the demographic transition, population growth increased along with investment in human capital, whereas the demographic transition brought about a decline in population growth along with a further increase in human capital formation.

Other unified theories examine several reinforcing mechanisms that could have triggered the demographic transition and the transition to sustained economic growth, such as the decline in child labor (Hazan and Berdugo 2002, Doepke 2004 and Doepke and Zilibotti 2003), the decline in mortality rates and the rise in life expectancy (Lagerlof 2003a, and Weisdorf 2004), and the evolution of preferences for offspring quality (Galor and Moav 2002)), as discussed in section 4.1. The quantitative unified theories of Fernandez-Villaverde (2003) and Doepke (2004) confirm the significance of these various channels in originating the demographic transition and the shift from stagnation to growth.

The Decline in the Gender Gap

The observed decline in the gender gap in the last two centuries, as discussed in section 3.3.4, is an alternative mechanism that could have triggered a demographic transition and human capital formation in other unified theories.

A unified theory based upon the decline in the gender wage gap and the associated increase in female labor force participation and fertility decline was explored by Galor and Weil (1996, 1999), as elaborated in section 3.3.4. They argue that technological progress and capital accumulation complemented mental intensive tasks and substituted for physical-intensive tasks in the industrial production process. In light of the comparative physiological advantage of men in physical-intensive tasks and women in mental-intensive tasks, the demand for women's labor input gradually increased in the industrial sector, decreasing monotonically the wage deferential between men and women. In early stages of

[117] Bloom and Williamson (1998) suggest that the cohort effect played a significant role in the growth "miracle" of East Asian countries in the time period 1960-1990.

[118] As established in section 3.3, some mechanisms that were proposed for the demographic transition, such as the decline in infant and child mortality, as well as the rise in income, are inconsistent with the evidence. These mechanisms were excluded in the formulation of unified growth theory.

industrialization, wages of men and women increased, but the rise in female's relative wages was insufficient to induce a significant increase in women's labor force participation. Fertility, therefore increased due to the income effect that was generated by the rise in men's absolute wages. Ultimately, however, the rise in women's relative wages was sufficient to induce a significant increase in labor force participation, increasing the cost of child rearing proportionally more that households income and triggering a demographic transition and a shift from stagnation to growth.

Similarly, a transition from stagnation to growth based upon a declining gender gap in human capital formation was proposed by Lagerlof (2003b). He argues that the process of development permitted a gradual improvement in the relative level of female education, raising the opportunity cost of children and initiating a fertility decline.[119]

4.2.3 Alternative Modeling of the Transition from Agricultural to Industrial Economy

The shift from agriculture to industry that accompanied the transition from stagnation to growth, as described in section 2.2.3, influenced the specifications of the production structure of most unified theories of economic growth. In some unified theories (e.g., Galor and Weil (2000)) the structure of the aggregate production function and its interaction with technological progress, reflects implicitly a transition from an agricultural to an industry economy in the process of development. In other theories (e.g., Hansen and Prescott 2002, Kogel and Prskawetz 2001, Hazan and Berdugo 2002, Tamura 2002, Doepke 2004, Galor and Mountford 2003, Bertocchi 2003, and Galor, Moav and Vollrath 2003) the process of development generates explicitly a transition from an agricultural sector to an industrial sector.

In Galor and Weil (2000) production occurs according to a constant-returns-to-scale technology that is subject to endogenous technological progress. The output produced at time t, is $Y_t = H_t^{\alpha}(A_t X)^{1-\alpha}$, where H_t is the aggregate quantity of efficiency units of labor employed in period t, X is land employed in production in every period t, and A_t represents the endogenously determined technological level in period t. Hence $A_t X$ are the "effective resources" employed in production in period t. In early stages of development, the economy is agricultural (i.e., the fixed amount of land is a binding constraint on the expansion of the economy). Population growth reduces labor productivity since the rate of technological progress is not sufficiently high to compensate for the land constraint. However as the rate of technological progress intensifies in the process of development the economy becomes industrial. Technological progress counterbalanced the land constraint, the role of land gradually diminishes, and "effective resources" are expanding at a rate that permit sustained economic growth.

Hansen and Prescott (2002) develop a model that captures explicitly the shift from an agricultural sector to an industrial sector in the transition from stagnation to growth. In early stages of development, the industrial technology is not sufficiently productive and production takes place solely in an agricultural sector, where population growth (that is *assumed* to increase with income) offset increases in productivity. An *exogenous* technological progress in the latent industrial technology ultimately makes the industrial sector economically viable and the economy gradually shifts resources from the agricultural sector to the industrial one. Assuming that the positive effect of income on population is reversed, the rise in productivity in the industrial sector is not counterbalanced by population growth permitting the transition to a state of sustained economic growth.

[119] Alternatively, one could have adopted the mechanism proposed by Fernandez, Fogli and Olivetti (2004) for the gradual decline in the education gap and labor force participation between men and women. They suggest that it reflects a dynamic process in which the home experience of sons of working, educated mothers makes them more likely to prefer educated and working wives, inducing a gradual increase in investment in education as well as labor force participation among women.

Unlike most unified theories in which the time paths of technological progress, population growth, and human capital formation are endogenously determined and are determined on the basis on explicit micro-foundations, in Hansen and Prescott (2002) technological progress is exogenous, population growth is assumed to follow the hump-shaped pattern that is observed along human history and human capital formation that appears central in the transition is absent. Based upon this reduced form approach, they demonstrate that there exists a rate of technological progress in the latent industrial sector and a well specified reduced form relationship between population and output under which the economy will shift from Malthusian stagnation to sustained economic growth. Unfortunately, however, this methodology does not advance us in identifying the underlying micro-foundations that led to the transition from stagnation to growth - the ultimate goal of unified growth theory.

Formally, the transition from stagnation to growth in Hansen and Prescott (2002) does not rely on the forces of human capital in the transition. However, the lack of a role for human capital in their structure is an artifact of the reduced form analysis that does not identify the economic factors behind the process of technological change in the latent industrial technology, as well as the forces behind the assumed hump shaped pattern of population dynamics. If the micro-foundations of these critical factors behind the transition would have been properly established, human capital would have played a central role sustaining the rate of technological progress in the industrial sector and in generating the demographic transition that is assumed in their setting.

Thus, no major insight has been generated from this explicit modeling of the transition from agriculture to industry in a closed economy setting. In contrast, the two-sector framework is instrumental in the exploring of the effect of international trade on the differential timing of the transition from stagnation to growth and the associate phenomenon of the great divergence (Galor and Mountford 2003), as discussed in section 6.1. Moreover, this two-sector setting would be necessary in order to examine the incentives of land owners to block education reforms and the process of industrialization (Galor, Moav, and Vollrath 2003), and the evolution of property rights and their impact on political reforms (Bertocchi 2003).

5 Unified Evolutionary Growth Theory

"It is not the strongest of the species that survive, nor the most intelligent, but the one most responsive to change." Charles Darwin

5.1 Human Evolution and Economic Development

This section explores the dynamic interaction between human evolution and the process of economic development. It focuses on a recent development of a unified evolutionary growth theory that, based on historical evidence, generates innovative hypotheses about the interplay between the process of development and human evolution, shedding new light about the origin of modern economic growth and the observed intricate evolution of health, life expectancy, human capital, and population growth since the Neolithic revolution.

The unified evolutionary growth theory advances a novel analytical methodology that is designed to capture the complexity of the dynamic interaction between the economic, social, and behavioral aspects of the process of development and evolutionary processes in the human population. The proposed hybrid between Darwinian methodology and the methodology of unified theories of economic growth permits the exploration of the dynamic reciprocal interaction between the evolution of the distribution of genetic traits and the process of economic development. It captures potential non-monotonic

evolutionary processes that were triggered by major socioeconomic transitions, and may have played a significant role in the observed time path of health, life expectancy, human capital, and population growth.[120]

Humans were subjected to persistent struggle for existence for most of their history. The Malthusian pressure affected the size of the population (as established in section 2.1.2), and conceivably via natural selection, the composition of the population as well. Lineages of individuals whose traits were complementary to the economic environment generated higher income, a larger number of surviving offspring and the representation of their traits in the population gradually increased, contributing significantly to the process of development.

Consistent with the historical evidence presented in section 2, the proposed unified evolutionary growth theory demonstrates that the Malthusian epoch that characterized most of human existence stimulated a process of natural selection that generated an evolutionary advantage to human traits that were complementary to the growth process, and ultimately generating a take-off from an epoch of Malthusian stagnation to a state of sustained economic growth.

Evidence suggests that evolutionary processes in the composition of existing genetic traits may be rather rapid and the time period between the Neolithic Revolution and the Industrial Revolution that lasted about 10,000 years is sufficient for significant evolutionary changes. There are numerous examples of rapid evolutionary changes among various species.[121] In particular, evidence establishes that evolutionary changes occurred in the Homo sapiens within the time period that is the focus of our analysis. For instance, lactose tolerance was developed among European and Near Easterners since the domestication of dairy animals in the course of the Neolithic revolution, whereas in regions that were exposed to dairy animals in later stages a larger proportion of the adult population suffers from lactose intolerance. Furthermore, genetic immunity to malaria provided by the sickle cell trait is prevalent among descendents of Africans whose engagement in agriculture improved the breeding ground for mosquitoes and thereby raised the incidence of malaria, whereas this trait is absent among descendents of nearby populations that have not made the transition to agriculture.[122]

Despite the existence of compelling evidence about the interaction between human evolution and the process of economic development, only few attempts have been made to explore the reciprocal interaction between the process of development and human evolution.[123] This exploration is likely to revolutionize our understanding of the process of economic development as well as the process of human

[120]The conventional methodology of *evolutionary stable strategies* that has been employed in various fields of economics, ignores the dynamics of the evolutionary process, and is thus inappropriate for the understanding of the "short-run" interaction between human evolution and the process of development since the Neolithic revolution. As will become apparent the dynamics of the evolutionary process are essential for the understanding of the interaction between human evolution and economic growth since the Neolithic revolution that was marked by fundamental non-monotonic evolutionary processes.

[121]The color change that peppered moths underwent during the 19th century is a classic example of evolution in nature [See Kettlewell 1973]. Before the Industrial Revolution light-colored English peppered moths blended with the lichen-covered bark of trees. By the end of the 19th century a black variant of the moth, first recorded in 1848, became far more prevalent than the lighter varieties in areas in which industrial carbon removed the lichen and changed the background color. Hence, a significant evolutionary change occurred within a time period that correspond to only hundreds of generations. Moreover, evidence from Daphne Major in the Galapagos suggests that significant evolutionary changes in the distribution of traits among Darwin's Finches occurred within few generations due to a major drought [Grant and Grant 1989]. Other evidence, including the dramatic changes in the color patterns of guppies within 15 generations due to changes in the population of predators, are surveyed by Endler [1986].

[122]See Levingston [1958], Weisenfeld [1967] and Durham [1982].

[123]Notable exceptions are Galor and Moav (2002)'s exploration of the interaction between human evolution and the transition from stagnation to growth, Saint Paul (2003)'s examination of the effect of the emergence of markets on the evolution of heterogeneity in the human population , Clark and Hamilton (2003) analysis of the relationship between the evolution of time preference and the sharp decline in interest rate in England in the 14th and 15th centuries, and Galor and Moav (2004)'s exploration of the effect of an increased population density in the process of development on the evolution of life expectancy.

evolution.

5.2 Natural Selection and the Origin of Economic Growth

The first evolutionary growth theory that captures the interplay between human evolution and the process of economic development in various phases of development, was developed by Galor and Moav (2002). The theory suggests that during the epoch of Malthusian stagnation that had characterized most of human existence, traits of higher valuation for offspring quality generated an evolutionary advantage and their representation in the population gradually increased. This selection process and its effect on investment in human capital stimulated technological progress and ultimately initiated a reinforcing interaction between investment in human capital and technological progress that brought about the demographic transition and the state of sustained economic growth.[124]

The theory maintains that during the Malthusian epoch, the distribution of valuation for quality lagged behind the evolutionary optimal level. The evolution of the human brain in the transition to Homo sapiens and the complementarity between brain capacity and the reward for human capital has increased the evolutionary optimal investment in the quality of offspring (i.e., the level that maximizes reproduction success).[125] Moreover, the increase in the return to human capital in the aftermath of the Neolithic revolution increased the evolutionary optimal levels of investment in child quality. The agricultural revolution facilitated the division of labor and fostered trade relationships across individuals and communities, enhancing the complexity of human interaction and raising the return to human capital. Thus, individuals with traits of higher valuation for offspring's quality generated higher income and, in the Malthusian epoch when child rearing was positively affected by aggregate resources, a larger number of offspring. Traits of higher valuation for quality gained the evolutionary advantage and their representation in the population increased over time.

The Malthusian pressure increased the representation of individuals whose preferences are biased towards child quality, positively affecting investment in human capital and ultimately the rate of technological progress. In early stages of development, the proportion of individuals with higher valuation for quality was relatively low, investment in human capital was minimal, resources above subsistence were devoted primarily to child rearing, and the rate of technological progress was rather slow. Technological progress therefore generated proportional increases in output and population and the economy was in the vicinity of a Malthusian equilibrium, where income per capita is constant, but the proportion of individuals with high valuation for quality was growing over time.[126]

As the fraction of individuals with high valuation for quality continued to increase, technological progress intensified, raising the rate of return to human capital. The increase in the rate of technological progress generated two effects on the size and the quality of the population. On the one hand, improved

[124]The theory is applicable for either social or genetic intergenerational transmission of traits. A cultural transmission is likely to be more rapid and may govern some of the observed differences in fertility rates across regions. The interaction between cultural and genetic evolution is explored by Boyd and Richardson (1985) and Cavalli-Sforza and Feldman (1981), and a cultural transmission of preferences is examined by Bisin and Verdier (2000).

[125]The evolutionary process in valuation for quality that was triggered by the evolution of the human brain has not reached a new evolutionary stable state prior to the Neolithic period because of the equality that characterized resource allocation among hunter-gatherers tribes. Given this tribal structure, a latent attribute of preferences for quality, unlike observable attributes such as strength and intelligence, could not generate a disproportionate access to sexual mates and resources that could affect fertility rates and investment in offspring's quality, delaying the manifestation of the potential evolutionary advantage of these traits. It was the emergence of the nuclear family in the aftermath of the agricultural revolution that fostered intergenerational links, and thereby enhanced the manifestation of the potential evolutionary advantage of this trait.

[126]Unlike Galor and Weil (2000) in which the adverse effect of limited resources on population growth delays the process of development, in the proposed theory the Malthusian constraint generates the necessary evolutionary pressure for the ultimate take-off.

technology eased households' budget constraints and provided more resources for quality as well as quantity of children. On the other hand, it induced a reallocation of these increased resources toward child quality. In the early stages of the transition from the Malthusian regime, the effect of technological progress on parental income dominated, and the rate of population growth as well as the average quality increased, further accelerating technological progress. Ultimately, the rate of technological progress induced universal investment in human capital along with a reduction in fertility rates, generating a demographic transition in which the rate of population growth declined along with an increase in the average level of education. The positive feedback between technological progress and the level of education reinforced the growth process, setting the stage for the transition to a state of sustained economic growth.[127]

During the transition from the Malthusian epoch to the sustained growth regime, once the economic environment improved sufficiently, the significance of quality for survival (fertility) declined, and traits of higher valuation for quantity gained the evolutionary advantage. Namely, as technological progress brought about an increase in income, the Malthusian pressure relaxed and the domination of wealth in fertility decisions diminished. The inherent advantage of higher valuation for quantity in reproduction has started to dominate, and individuals whose preferences are biased towards child quantity gained the evolutionary advantage. Nevertheless, the growth rate of output per worker has remained positive since the high rate of technological progress sustained an attractive return to investment in human capital even from the viewpoint of individuals whose valuation for quality is relatively low.

The transition from stagnation to growth is an inevitable by product of the interaction between the composition of the population and the rate of technological progress in the Malthusian epoch. However, for a given composition of population, the timing of the transition may differ significantly across countries and regions due to historical accidents, as well as variation in geographical, cultural, social and institutional factors, trade patterns, colonial status, and public policy, that have affected the relationship between human capital formation and technological progress.

5.2.1 Primary Ingredients

The theory is based upon the interaction between several building blocks: the Darwinian elements, the Malthusians elements, the nature of technological progress, the determinants of human capital formation, and the factors that affect parental choice regarding the quantity and quality of offspring.

The Darwinian elements. The theory incorporates the main ingredients of Darwinian evolution (i.e., variety, intergenerational transmission of traits, and natural selection) into the economic environment. Inspired by fundamental components of the Darwinian theory (Darwin 1859, 1871), individuals do not operate consciously so as to assure their evolutionary advantage. Nevertheless, their preferences (or strategies) assure that those individuals whose operations are most complementary to the environment would ultimately dominate the population.

Individuals' preferences are defined over consumption above a subsistence level as well as over the quality and the quantity of their children.[128] These preferences capture the Darwinian survival

[127]The theory suggests that waves of rapid technological progress in the Pre-Industrial Revolution era did not generate sustained economic growth due to the shortage of preferences for quality in the population. Although technological progress increased the return to quality temporarily, in these previous episodes, the level of human capital that was generated by the response of the existing population was insufficient to sustain technological progress and economic growth.

[128]The subsistence consumption constraint is designed to capture the fact that the physiological survival of the parent is a pre-condition for the survival of the lineage (dynasty). Resources allocated to parental consumption beyond the subsistence level may be viewed as a force that raises parental productivity and resistance to adverse shocks (e.g., famine and disease), generating a positive effect on the fitness of the parent and the survival of the lineage. This positive effect, however, is counterbalanced by the implied reduction in resources allocated to the offspring, generating a negative effect on the survival of the lineage.

strategy as well as the most fundamental trade-offs that exist in nature: namely, the trade-off between the resources allocated to the parent and the offspring, and the trade-off between the number of offspring and resources allocated to each offspring.[129] The economy consists of a variety of types of individuals distinguished by the weight given to child quality in their preferences.[130] This trait is assumed to be transmitted intergenerationally, the economic environment determines the type with the evolutionary advantage (i.e., the type characterized by higher fertility rates), and the distribution of preferences in the population evolves over time due to differences in fertility rates across types.[131]

The significance that individuals attribute to child quantity as well as to child quality reflects the well-known variety in the quality-quantity survival strategies (or in the K and r strategies) that exists in nature (e.g., MacArthur and Wilson 1967). Human beings, like other species, confront the basic trade-off between offspring's quality and quantity in their implicit Darwinian survival strategies. Although a quantity-biased preference has a positive effect on fertility rates and may therefore generate a direct evolutionary advantage, it adversely affects the quality of offspring, their income, and their fitness and may therefore generate an evolutionary disadvantage. "Increased bearing is bound to be paid for by less efficient caring." (Dawkins 1989, p. 116). As was established in the evolutionary biology literature since the seminal work of Lack (1954), the allocation of resources between offspring "caring" and "bearing" is subjected to evolutionary changes.[132]

The Malthusian elements. Individuals are subjected to a subsistence consumption constraint and as long as the constraint is binding, an increase in income results in an increase in population growth along with an increase in the average quality of a minor segment of the population. Technological progress, which brings about temporary gains in income per capita, triggered therefore in early stages of development an increase in the size of the population that offset the gain in income per capita due to the existence of diminishing returns to labor. Growth in income per capita is generated ultimately, despite decreasing returns to labor, since technological progress induces investment in human capital among a growing minority.

The determinants of technological progress. The composition of the population as reflected by the average level of human capital is the prime engine of technological progress.[133]

[129]Resources allocated to quality of offspring in different stages of development take different forms. In early stages of development it is manifested in investment in the durability of the offspring via better nourishment and parental guidance, whereas in mature stages, investment in quality may capture formal education.

[130]The analysis abstracts from heterogeneity in the degree of the trade-off between resources allocated to parent and offspring. The introduction of this element would not alter the qualitative results. On the evolution of preferences see the survey by Bowles (1998).

[131]Recent research across historical and modern data from the United States and Europe suggests that fertility behavior has a significant hereditary component [Rogdgers et al. 2001a]. For instance, as established recently by Kohler et al. (1999) and Rodgers et al. (2001b), based on the comparison of fertility rates among identical and fraternal twins born in Denmark during the periods 1870-1910 and 1953-1964, slightly more than one-quarter of the variance in completed fertility is attributable to genetic influence. These findings are consistent with those of Rodgers and Doughty [2000] based on kinship data from the United States.

[132]Lack (1954) suggests that clutch sizes (i.e., number of eggs per nest), among owls and other predatory vole-eating birds, for instance, are positively related to food abundance. He argues that the clutch size is selected such that under any feeding conditions fertility rates ensure the maximal reproductive success. Furthermore, Cody (1966) documents the existence of significant differences between clutch sizes of the same bird species on islands and nearby mainland localities of the same latitude. In temperate regions where food is more abundant in the mainland than on islands, the average clutch size is smaller on the islands. For instance, for *Cyanoramphus novaezelandeae*, the average mainland clutch is 6.5 whereas the average in the island is 4.

[133]This link between education and technological change was proposed by Nelson and Phelps (1966) and was supported empirically by Easterlin (1981), Doms et al. (1997), as well as others. Consistently with Mokyr (2002) who argues that the effect of human capital accumulation on technological progress became significant only in the course of the Scientific Revolution that preceded the Industrial Revolution, the effect of human capital accumulation on the rate of technological progress, need not be significant prior to the scientific revolution as long as it becomes significant prior to the Industrial Revolution. In order to focus on the role of the evolutionary process, the model abstracts from the potential positive effect of the size of the population on the rate of technological progress. Adding this scale effect would simply accelerate the

The origin of human capital formation. Technological change raises the demand for human capital. Technological progress reduces the adaptability of existing human capital for the new technological environment and educated individuals (and thus offspring of parent with high valuation for quality) have a comparative advantage in adapting to the new technological environment.[134]

The determination of paternal decision regarding offspring quantity and quality. Individuals choose the number of children and their quality based upon their preferences for quality as well as their time constraint.[135] The rise in the (genetic or cultural) bias towards quality in the population, as well as the rise in the demand for human capital, induce parents to substitute quality for quantity of children.[136]

5.2.2 Main Hypotheses and their Empirical Assessment

The theory generates several hypotheses about human evolution and the process of development, underlying the role of natural selection in: (i) the gradual process of human capital formation and thus technological progress prior to the Industrial Revolution, and (ii) the acceleration of the interaction between human capital and technological progress in the second phase of the Industrial Revolution, the associated demographic transition, and the emergence of a state of sustained economic growth.

The Main Hypotheses:

- During the initial phases of the Malthusian epoch, the growth rate of output per capita is nearly zero and the growth rate of population and literacy rates is minuscule, reflecting the sluggish pace of technological progress, the low representation of individuals with high valuation for child quality, and the slow pace of the evolutionary process.

This hypothesis is consistent with the characteristics of the Malthusian epoch, as described in section 2.1.

- In the pre-demographic transition era, traits for higher valuation for offspring quality generated an evolutionary advantage. Namely, individuals with higher valuation for the quality of children had a larger number of surviving offspring and their representation in the population increased over time. In contrast, in the post-demographic transition era, when income per capita has no longer been the binding constraint on fertility decisions, individuals with higher valuation for offspring quantity have had an evolutionary advantage, bearing a lager number of surviving offspring. Thus, in the pre-demographic transition era, the number of surviving offspring was affected positively by parental education and parental income whereas in the post-demographic transition era, in contrast, this pattern is reversed and more educated, higher income individuals have a smaller number of surviving offspring.

transition process (e.g., Galor and Weil 2000).

[134] See Schultz (1964) and Nelson and Phelps (1966). If the return to education rises with the level of technology rather than with the rate of technological progress, the qualitative analysis would not be affected. However, this alternative would imply that changes in technology were skill-biased throughout human history in contrast to those periods in which technological change was skilled-saving, notably, in the first phase of the Industrial Revolution.

[135] Anthropological evidence suggests that fertility control was indeed exercised even prior to the Neolithic Revolution. Reproductive control in hunter-gatherer societies is exemplified by "pacing birth" (e.g., birth every four years) conducted by tribes who live in small, semi nomadic bands in Africa, Southeast Asia, and New Guinea in order to prevent the burden of carrying several children while wandering. They abstained from sexual intercourse for a three-year period after each birth. Similarly, Nomadic women of the Kung (a group of the San people of Southern Africa), use no contraceptives but nurse their babies frequently, suppressing ovulation and menstruation for two to three years after birth, and reaching a mean interval between births of 44 months.

[136] The existence of a trade-off between quantity and quality of children is supported empirically (e.g., Rosenzweig and Wolpin (1980) and Hanushek (1992)).

Clark and Hamilton (2003) examine empirically this hypothesis on the basis of data that they have constructed from wills written in England in the time period 1620-1636. The wills that were written in a closed proximity to the death of a person in urban and rural areas, across a large variety of occupations and wealth, contain information about the number of surviving offspring, literacy of testator (measured by whether the will was signed), occupation of testator (if male), the amount of money bequeathed and to whom (spouse, children, the poor, unrelated persons), and houses and land that were bequeathed. Based on this data, Clark and Hamilton find a positive and statistically significant effect of literacy (and wealth) on the number of surviving offspring.[137] They confirm the hypothesis that literate people (born, according to the theory, to parents with quality-bias) had an evolutionary advantage in this (pre-demographic transition) period.[138] The negative relationship between education and fertility within a country in the post-demographic transition era was documented extensively.[139]

- The process of natural selection prior to the Industrial Revolution increased the representation of individuals with higher valuation for quality, gradually increasing the average level of investment in human capital,[140] permitting a slow growth of output per capita.

The prediction about the rise in literacy rates prior to the Industrial Revolution is consistent with historical evidence. Various measures of literacy rates demonstrate a significant rise in literacy rates in the 2 centuries that preceded the Industrial Revolution in England.[141] As depicted in Figure 5.1, male literacy rates increased gradually in the time period 1600-1760. Literacy rates for men doubled over this period, rising from about 30% in 1600 to over 60% in 1760. Similarly, as reported by Cipolla (1968), literacy rates of women more than tripled from less than 10% in 1640 to over 30% in 1760.[142]

[137] In addition, Boyer (1989) argues that in early nineteenth century England, agricultural laborers' income had a positive effect on fertility: birth rates increased by 4.4 percent in response to a 10 percent increase in annual income. Further evidence are surveyed by Lee (1997).

[138] Interestingly, in New France, where land was abundant, and thus fertility decisions were not constraint by the availability of resources, the number of surviving offspring was higher among less educated individuals. These findings are consistent with the theory as well. If resource constraint is not binding for fertility decisions (e.g., in the post-demographic transition era, or due to a positive shock to income in the Malthusian era), individuals with higher valuation for quantity gain an evolutionary advantage.

[139] See, for instance, Kremer and Chen (2002).

[140] In contrast to Galor and Weil (2000) in which the inherent positive interaction between population and technology during the Malthusian regime is the force behind the increase in the rate of technological progress that induced investments in human capital and led to further technological progress, a demographic transition, and sustained economic growth, Galor and Moav (2002) is structured such that the gradual change in the composition of the population (rather than by the size of the population) brings about the take-off from Stagnation to growth. Thus, a scale effect is not needed for the take-off. However, this is just a simplifying modelling devise and both forces could operated simultaneously in triggering the take-off.

[141] Moreover, this hypothesis appears consistent with the increase in the number and size of universities in Europe since the establishment of the first university in Bologna in the 11th century, significantly outpacing the growth rate of population.

[142] This pattern is robust and is observed in various diocese over this period. For instance Cressy (1981, table 6.3 p. 113) reports a gradual rise in average literacy rate of average of yeomen, husbandmen and tradesmen in Norwich from 30% in 1580 to nearly 61% in 1690, and Cressy (1980, table 7.1 p. 143) reports a gradual rise in Gentle literacy in the diocese of Durham over the period 1565 to 1624.

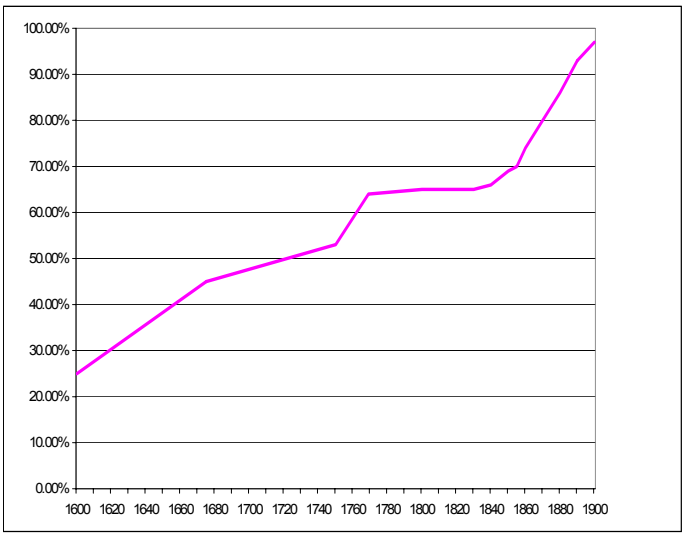


Figure 5.1 The Rise in Male Literacy Rates Prior and During the Industrial Revolution: England: 1600-1900
Sources: Cipolla (1969), Stone (1969) and Schofield (1973)

Moreover, as argued by Clark (2003), human capital accumulation in England began in an era when the market rewards to skill acquisition were at historically low levels, consistent with the argument that the rise in human capital reflected a rise in the preference for quality offspring.

- The acceleration in the rate of technological progress that was reinforced by the investment in human capital of individuals with high valuation for offspring quality, increased the demand for human capital in the Post-Malthusian Regime, generating a universal investment in human capital, a demographic transition and a rapid pace of economic growth.

The hypothesis is consistent with the evidence, provided in section 2.3 and depicted partly in Figure 4.1, about the significant rise in the demand for human capital in the second phase of the Industrial Revolution, the marked increased in educational attainment, and the emergence of universal education towards the end of the 19th century in association with a decline in fertility rates, and a transition to a state of sustained economic growth.

5.3 Complementary Mechanisms

The theory argues that during the Malthusian epoch hereditary human traits, physical or mental, that generate higher earning capacity, and thereby potentially larger number of offspring, would generate an evolutionary advantage and would dominate the population in the long run. Hereditary traits that stimulate technological progress or raise the incentive to invest in offspring's human capital (e.g., ability, longevity, and a preference for quality), may trigger a positive feedback loop between investment in human capital and technological progress that would bring about a take-off from an epoch of Malthusian stagnation, a demographic transition and a shift to a state of sustained economic growth. Hence, the struggle for existence that had characterized most of human history stimulated natural selection and generated an evolutionary advantage to individuals whose characteristics are complementary to the growth process, ultimately triggering a take-off from an epoch of stagnation to sustained economic growth. Galor and Moav (2002) focus on the evolution of the trade-off between resources allocated to

the quantity and the quality of offspring. Their framework of analysis can be modified to account for the interaction between economic growth and the evolution of other hereditary traits.

5.3.1 The Evolution of Ability and Economic Growth

Consider the model described earlier. Suppose that individual's preferences, are defined over consumption above a subsistent level and over child quality and quantity. Individuals are identical in their preferences, but differ in their hereditary innate ability. Suppose further that offspring's level of human capital is an increasing function of two complementary factors: innate ability and investment in quality. Thus, since the marginal return to investment in child quality increases with ability, higher ability individuals and hence dynasties would allocate a higher fraction of their resources to child quality.

In the Malthusian era individuals with a higher ability generate more income and hence are able to allocate more resources for child quality and quantity. High ability individuals, therefore, generate higher income due to fact that their innate ability as well as their quality are higher. In the Malthusian era fertility rates are positively affected by the level of income and (under plausible configurations) the high ability individuals have therefore an evolutionary advantage over individuals of lower ability. As the fraction of individuals of the high ability type increases, investment in quality rises, and technological progress intensified. Ultimately the dynamical system changes qualitatively, the Malthusian temporary steady-state vanishes endogenously and the economy takes-off from the Malthusian trap. Once the evolutionary process generates the positive feedback between the rate of technological progress and the level of education, technological progress is reinforced, the return to human capital increases further, setting the stage for the demographic transition and sustained economic growth.

5.3.2 The Evolution of Life Expectancy and Economic Growth

Suppose that individuals differ in their level of health due to hereditary factors. Suppose further that there exist a positive interaction between the level of health and economic well-being. Higher income generates a higher level of health, whereas higher level of health increases labor productivity and life expectancy. Parents that are characterized by high life expectancy and thereby expect their offspring to have a longer productive life, would allocate more resources toward child quality. In the Malthusian era fertility rates are positively affected by the level of income and individuals with higher life expectancy, and therefore higher quality and higher income, would have (under plausible configurations) an evolutionary advantage. Natural selection therefore, increases the level of health as well as the quality of the population. Eventually, this process generates a positive feedback loop between investment in child quality, technological progress and health, bringing about a transition to sustained economic growth with low fertility rates and high longevity.

Alternatively, Galor and Moav (2004b) hypothesize that major socioeconomic and environmental changes in the process of development that were associated with significant increases in population density (e.g., the Neolithic Revolution and the process of urbanization) triggered evolutionary processes that contributed significantly to the long lasting improvements in longevity. Consistent with historical evidence, the theory suggests that the rise in the extrinsic mortality rate, brought about by a denser population, led initially to a decline in life expectancy, as document for instance in Figure 2.8.[143] However, the evolutionary process that was originated by the rise in mortality, gradually increased the representation of traits associated with resistance to diseases and thus higher life expectancy, contributing

[143]Moreover, the evolutionary forces that major epidemics (e.g., the Black Death) and climatic changes have triggered, might have influenced the time path of human longevity and economic development in the subsequent centuries.

significantly to the observed rise in longevity. Moreover, this evolutionary process in life expectancy reinforced the interaction between investment in human capital, life expectancy, and technological progress thereby expediting the demographic transition and enhancing the economic transition from stagnation to growth.[144]

5.4 Assessment of the Various Mechanisms

The significance of the evolution of various genetic traits in the transition from an epoch of Malthusian stagnation to a state of sustained economic growth, ought to consider the possibility that some of these traits may have completed most of their evolutionary change tens of thousands of years before the take-off and may be therefore a pre-condition for the take-off rather than the trigger itself.

In particular, the conventional wisdom among evolutionary biologists is that intelligence has not evolved markedly since the emergence of Homo Sapience (i.e., intelligence may have reached a temporary evolutionary optimum, reflecting the trade-off between the benefits and the energy cost associated with a lager brain). In contrast, it is unlikely that preferences reflecting quality-bias would have reached a an evolutionary stable state very early in the evolution of mankind. Prior to the Neolithic period, the majority of people lived in tribes where resources as well as child rearing were shared by the community. Given this tribal structure, the latent attribute of preferences for quality, unlike observable attributes such as strength and intelligence, could not generate a disproportionate access to sexual mates and resources that could affect fertility rates and investment in offspring's quality, delaying the manifestation of the potential evolutionary advantage of individuals with a quality-bias. It was the emergence of the nuclear family in the aftermath of the agricultural revolution that fostered intergenerational links, and thereby enhanced the manifestation of the potential evolutionary advantage of individuals with a quality-bias.[145]

6 Differential Takeoffs and the Great Divergence

The last two centuries have witnessed dramatic changes in the distribution of income and population across the globe. The differential timing of the take-off from stagnation to growth across countries and the corresponding variations in the timing of the demographic transition have led to a great divergence in income as depicted in Figure 2.32 and to significant changes in the distribution of population around the globe, as depicted in Figure 2.33. Some regions have excelled in the growth of income per capita, while other regions have been dominant in population growth.[146]

Inequality in the world economy had been insignificant until the 19th century. The ratio of GDP per capita between the richest region and the poorest region in the world was only 1.1:1 in the year 1000, 2:1 in the year 1500 and 3:1 in the year 1820. In contrast, the past two centuries have been characterized

[144]The evolution of the human brain along with the evolution of life expectancy in the prior to the Neolithic revolution is examined by Robson and Kaplan (2003).

[145]An alternative explanation for the delay in the evolutionary process of the quality bias relative to the evolution of ability is based on the notion of punctuated equilibria (Gould 1977). A sequence of mutations, which result in a gradual increase in the variance in the distribution of the (latent) quality bias trait, had not affected investment in offspring's quality for a long period due to the low rate of return to human capital. Ultimately, however, mutations increased the variance sufficiently so as to induce investment in offspring's quality, despite the low return, and brought about an evolutionary advantage for the quality type. In contrast, a gradual increase in the variance of non-latent variables, such as ability, would have an immediate effect on the evolutionary process.

[146]Some researchers (e.g., Jones (1997) and Pritchett (1997)) have demonstrated that the great divergence that has been witnessed in the last two centuries has been maintained in the last decades as well, across countries. Interestingly, however, as established by Sala-i-Martin (2002), the phenomena has not been maintained across people in the world, (i.e., when national boundaries are removed).

by a 'Great Divergence' in income per capita among countries and regions. In particular, the ratio of GDP per capita between the richest and the poorest regions has widened considerably from a modest 3:1 ratio in 1820, to a large 18:1 ratio in 2001. An equally impressive transformation occurred in the distribution of world population across regions, as depicted in Figure 2.33. The earlier take-off of Western European countries generated a 16% increase in the share of their population in the world economy within the time period 1820-1870. However, the early onset in the Western European demographic transition, and the long delay in the demographic transition of less developed regions well into the second half of the twentieth century, led to a 55% decline in the share of Western European population in the world in the time period 1870-1998. In contrast, the prolongation of the Post-Malthusian period of less developed regions and the delay in their demographic transition, generated a 84% increase in Africa's share of world population, from 7% in 1913 to 12.9% in 1998, an 11% increase in Asia's share of world population from 51.7% in 1913 to 57.4% in 1998, and a four-fold increase in Latin American's share in world population from 2% in 1820 to 8.6% in 1998.

The phenomenon of the Great Divergence in income per capita across regions of the world in the past two centuries, that was associated with the take-off from the epoch of near stagnation to a state of sustained economic growth, presents intriguing questions about the growth process. How does one account for the sudden take-off from stagnation to growth in some countries in the world and the persistent stagnation in others? Why has the positive link between income per capita and population growth reversed its course in some economies but not in others? Why have the differences in per capita incomes across countries increased so markedly in the last two centuries? Has the transition to a state of sustained economic growth in advanced economies adversely affected the process of development in less-developed economies?

6.1 Non-Unified Theories

The origin of the Great Divergence has been a source of controversy. The relative role of geographical and institutions factors, ethnic, linguistic, and religious fractionalization, colonialism and globalization has been in the center of a debate about the origins of this remarkable change in the world income distribution in the past two centuries.

The role of institutional and cultural factors has been the focus of influential hypotheses regarding the origin of the great divergence. North (1981), Landes (1998), Mokyr (1990, 2002), Hall and Jones (1999), Parente and Prescott (2000), and Acemoglu, Johnson and Robinson (2002) have argued that institutions that facilitated the protection of property rights and enhanced technological research and the diffusion of knowledge, have been the prime factors that enabled the earlier European take-off and the great technological divergence across the globe.[147]

The effect of geographical factors on economic growth and the great divergence have been emphasized by Jones (1981), Diamond (1997) and Gallup, Sachs, and Mellinger (1998).[148] The geographical hypothesis suggests that advantageous geographical conditions made Europe less vulnerable to the risk associated with climate and diseases, leading to the early European take-off, whereas adverse geographical conditions (e.g. harsh climate, prevalence of diseases, scarcity of natural resources, high transportation costs, limited regional diffusion of knowledge and technology) in disadvantageous regions, generated permanent hurdles for the process of development, contributing to the great divergence.[149]

[147]Barriers to technological adoption that may lead to divergence are explored by Caselli and Coleman (2002), Howitt and Mayer-Foulkes (2002) and Acemoglu, Aghion and Zilibotti (2003) as well.

[148]See Hall and Jones (1999), Masters and McMillan (2001) and Hibbs and Olson (2004) as well.

[149]Bloom, Canning and Sevilla (2003) cross section analysis rejects the geographical determinism, but maintain nevertheless that favorable geographical conditions have mattered for economic growth since they increase the likelihood of an

Recent research by Engerman and Sokoloff (2000) and Acemoglu, Johnson and Robinson (2002) propose that initial geographical conditions had a persistent effect on the quality of institutions, leading to divergence and overtaking in economic performance. Engerman and Sokolof (2000) provide descriptive evidence that geographical conditions that led to income inequality, brought about oppressive institutions designed to preserve the existing inequality, whereas geographical characteristics that generated an equal distribution of income led to the emergence of growth promoting institutions. Acemoglu, Johnson and Robinson (2002) provide evidence that reversals in economic performance across countries have a colonial origin, reflecting institutional reversals that were introduced by European colonialism across the globe.[150] "Reversals of fortune" reflect the imposition of extractive institutions by the European colonialists in regions in which favorable geographical conditions led to prosperity, and the implementation of growth enhancing institutions in poorer regions.[151]

Furthermore, the role of ethnic, linguistic, and religious fractionalization in the emergence of divergence and "growth tragedies" has been linked to their effect on the quality of institutions. Easterly and Levine (1997) and Alesina et al. (2003) demonstrate that geopolitical factors brought about a high degree of fractionalization in some regions of the world, leading to the implementation of institutions that are not conducive for economic growth and thereby to diverging growth paths across regions.

Empirical research suggests that indeed initial geographical conditions affected the current economic performance primarily via their effect on institutions. Acemoglu, Johnson and Robinson (2002), Easterly and Levine (2003), and Rodrik, Subramanian and Trebbi (2004) provide evidence that variations in the contemporary growth processes across countries can be attributed to institutional factors whereas geographical factors are secondary, operating primarily via variations in institutions.

A theory that unifies the geographical and the institutional paradigms, capturing the transition from the domination of the geographical factors in the determination of productivity in early stages of development to the domination of the institutional factors in mature stages of development has been proposed by Galor, Moav and Vollrath (2003). The theory identifies and establishes the empirical validity of a novel channel through which favorable geographical conditions that were inherently associated with inequality affected the emergence of human capital promoting institutions (e.g., public schooling, child labor regulations, abolishment of slavery, etc.), and thus the pace of the transition from an agricultural to an industrial society.[152] They suggest that the distribution of land within and across countries affected the nature of the transition from an agrarian to an industrial economy, generating diverging growth patterns across countries. The accumulation of physical capital in the process of industrialization has raised the importance of human capital in the growth process, reflecting the complementarity between capital and skills. Investment in human capital, however, has been sub-optimal due to credit markets imperfections, and public investment in education has been growth enhancing. Nevertheless, human capital accumulation has not benefited all sectors of the economy. Due to a low degree of complementarity between human capital and land, universal public education has increased the cost of labor beyond the increase in average labor productivity in the agricultural sector, reducing the return to land. Landowners, therefore, had no economic incentives to support these growth enhancing educational policies as long as their stake in the productivity of the industrial sector was insufficient. Land abundance, which was beneficial in early stages of development, brought about a hurdle for human capital accu-

economy to escape a poverty trap.

[150] Additional aspects of the role of colonialism in comparative development are analyzed by Bertocchi and Canova (2002).

[151] Brezis, Krugman and Tsiddon (1993) attribute technological leapfrogging to the acquired comparative advantage (via learning by doing) of the current technological leaders in the use of the existing technologies.

[152] As established by Chanda and Dalgaard (2003), variations in the structural composition of economies and in particular the allocation of scarce inputs between the agriculture and the non-agriculture sectors are important determinants of international differences in TFP, accounting for between 30 and 50 percents of these variations.

mulation and economic growth among countries that were marked by an unequal distribution of land ownership.[153]

6.2 A Unified Theory - Globalization and the Great Divergence

Unified theories of economic growth generate direct hypotheses about the factors that determine the timing of the transition from stagnation to growth and thus the factors that contributed to the Great Divergence. The timing of the transition may differ significantly across countries and regions due to historical accidents, as well as variation in geographical, cultural, social and institutional factors, trade patterns, colonial status, and public policy, that have affected the relationship between human capital formation and technological progress.[154]

This section explores a unified growth theory that generates a transition from stagnation to growth along with a great divergence, focusing on the asymmetric effect of globalization on the timing of the take-off from the Malthusian epoch of developed and less developed countries. Galor and Mountford (2003) suggest that sustained differences in income and population growth across countries may be attributed to the contrasting effect of international trade on industrial and non-industrial nations. Consistent with the evidence provided in section 2, their theory suggests that the expansion of international trade in the 19th century and its effect on the pace of individualization has played a major role in the timing of demographic transitions across countries and has thereby been a significant determinant of the distribution of world population and a prime cause of the 'Great Divergence' in income levels across countries in the last two centuries. International trade had an asymmetrical effect on the evolution of industrial and non-industrial economies**.** While in the industrial nations the gains from trade were directed primarily towards investment in education and growth in output per capita, a significant portion of the gains from trade in non-industrial nations was channeled towards population growth.[155]

In the second phase of the Industrial Revolution, international trade enhanced the specialization of industrial economies in the production of industrial, skilled intensive, goods. The associated rise in the demand for skilled labor has induced a gradual investment in the quality of the population, expediting a demographic transition, stimulating technological progress and further enhancing the comparative advantage of these industrial economies in the production of skilled intensive goods. In non-industrial economies, in contrast, international trade has generated an incentive to specialize in the production of unskilled intensive, non-industrial, goods. The absence of significant demand for human capital has provided limited incentives to invest in the quality of the population and the gains from trade have been utilized primarily for a further increase in the size of the population, rather than the income of the existing population. The demographic transition in these non-industrial economies has been significantly delayed, increasing further their relative abundance of unskilled labor, enhancing their comparative dis-

[153] An alternative mechanism is explored by Berdugo, Sadik and Sussman (2003).

[154] Related to the unified paradigm, Pomeranz (2000) has suggested that the discovery of the New World enabled Europe, via Atlantic trade, to overcome 'land constraints' and to take-off technologically. The inflow of grain and other commodities as well as the outflow of migrants during the Nineteenth century may have played a crucial role in Europe's development. By easing the land constraint at a crucial point — when income per capita had begun to rise rapidly, but before the demographic transition had gotten under way — the "ghost acres" of the New World provided a window of time which allowed Europe to pull decisively away from the Malthusian equilibrium.

[155] In contrast to the recent literature on the dynamics of comparative advantage (e.g., Findlay and Kierzkowski (1983), Grossman and Helpman (1991) Matsuyama (1991), Young (1991), Mountford (1998), and Baldwin et. al (2001) the focus on the interaction between population growth and comparative advantage and the persistent effect that this interaction may have on the distribution of population and income in the world economy generates an important new insight regarding the distribution of the gains from trade. The theory suggests that even if trade affects output growth of the trading countries at the same rate, (due to the terms of trade effect) income *per capita* of developed and less developed economies will diverge since in less developed economies growth of total output will be generated primarily by population growth, whereas in developed economies it will be generated by an increase in output per capita.

advantage in the production of skilled intensive goods and delaying their process of development. The research suggests, therefore, that international trade affected persistently the distribution of population, skills, and technologies in the world economy, and has been a significant force behind the 'Great Divergence' in income per capita across countries.[156]

The historical evidence described in section 2 suggests that the fundamental hypothesis of this theory is consistent with the process of development of the last two centuries. As implied by the trade patterns reported in Table 2.1, and the evolution of industrialization depicted in Figure 2.14, trade over this period induced the specialization of industrialized economies in the production of industrial goods whereas non-industrial economies specialize in the production of primary goods. The asymmetric effect of international trade on the process of industrialization of developed and less developed economies, as depicted in Figure 2.14, affected the demand for human capital as analyzed in section 2.3.3, and thus the timing of the demographic transition in developed and less-developed economies, generating a great divergence in output per capita as well as significant changes in the distribution of world population, as depicted in Figure 2.33.[157]

The diverging process of development of the UK and India since the 19th century in terms of the levels of income per capita and population growth is consistent with the theory of Galor and Mountford (2003) and provides an interesting case study. During the nineteenth century the UK traded manufactured goods for primary products with India.[158] Trade with Asia constituted over 20% of UK total exports and 23.2% of total imports throughout the nineteenth century (Bairoch 1974). [159] Consistent with the proposed hypothesis, as documented in Figure 2.14, industrialization in the UK accelerated, leading to a significant increase in the demand for skilled labor in the second phase of the Industrial Revolution, a demographic transition and a transition to a state of sustained economic growth.

For India, however, international trade played the reverse role. The period 1813-1850 was characterized by a rapid expansion in the volume of exports and imports which gradually transformed India from being an exporter of manufactured products – largely textiles – into a supplier of primary commodities (Chaudhuri 1983). Trade with the UK was fundamental in this process, with the UK supplying over two thirds of its imports for most of the nineteenth century and being the market for over a third of India's exports. As depicted in Figure 2.14, the rapid industrialization in the UK in the nineteenth century was associated with a decline in the per capita level of industrialization in India.[160] The delay in the process of industrialization and consequently the lack of demand for skilled labor delayed the de-

[156]Consistent with the thesis that human capital has reinforced the existing patterns of comparative advantage, Taylor (1999) argues that human capital accumulation during the late Nineteenth Century was not a source of convergence even among the advanced ' Greater Atlantic' trading economies. The richer economies - U.S.A. and Australia – had greater levels of school enrollments than the poorer ones, Denmark and Sweden.

[157]Consistent with the viewpoint the trade has not been uniformly beneficial across time and regions, recent research by Rodriguez and Rodrik (2001) has indicated that the relationship between openness and growth changed in the last century. Moreover, Clemens and Williamson (2004) find a positive relationship between average tariff levels and growth for the period 1870-1913 and a negative relationship for the period 1970-1998. Similarly Vamvakadis (2002) finds a positive relationship between several measures of openness and growth after 1970 and some evidence of a negative relationship in the period 1870-1910.

[158]The colonial power of the UK may have encouraged the specialization of India in the production of primary goods beyond the degree dictated by market forces. However, these forces would have just reinforced the adverse effects described in this paper.

[159]In contrast, trade with Asia constituted only 5% or less of French, German or Italian exports and 12.1% of total imports of continental Europe.

[160]Furthermore, Bairoch (1974) found that industries that employed new technologies made up between 60 and 70 percent of the UK manufacturing industry in 1860 but less than 1 percent of manufacturing industries in the developing countries. This contrasts with the experience of the non-UK European economies which produced more of the 'new technology' goods and which traded with themselves to a greater extent, (Bairoch, 1974).

mographic transition and the process of development.[161] Thus, while the gains from trade were utilized in the UK primarily towards an increase in output per capita, in India they were channeled towards an increase in the size of the population. The ratio of output per capita in the UK relative to India grew from 3:1 in 1820 to 11:1 in 1998, whereas the ratio of India's population relative to the UK's population grew from 8:1 in 1870 to 16:1 in 1998.[162]

7 Concluding Remarks

The transition from stagnation to growth and the associated phenomenon of the great divergence have been the subject of an intensive research in the growth literature in recent years. The discrepancy between of exogenous and endogenous neoclassical growth models and the process of development along most of human history, induced growth theorists to advance an alternative theory that captures in a single unified framework the contemporary era of sustained economic growth, the epoch of Malthusian stagnation that had characterized most of human history, and the fundamental driving forces of the recent transition between these distinct regimes.

The understanding of the contemporary growth process is fragile and incomplete unless growth theory would be based on proper micro-foundations that would reflect the qualitative aspects of the growth process in its entirety. Moreover, a comprehensive understanding of the hurdles faced by less developed economies in reaching a state of sustained economic growth would be futile unless the origin the transition of the currently developed economies into a state of sustained economic growth would be identified and their implications would be modified to account for the differences in the structure of less developed economies in an interdependent world.

Imposing the constraint that a single unified theory account for the entire intricate process of development in the last thousands of years is a discipline that enhances the viability of growth theory. A unified theory of economic growth reveals the underlying micro foundations that are consistent with the process of economic development along the entire spectrum of human history, rather than with the last century only, enhancing the confidence in the viability of growth theory, its predictions and policy implications, while improving the understanding of the sources of the recent transition from stagnation to growth and the associated phenomenon of the great divergence.

Unified growth theory suggests that the transition from stagnation to growth is an inevitable by-product of the inherent Malthusian interaction between population and technology and its ultimate impact on the demand for human capital and thereby on the onset of the demographic transition . Variations in the timing of the transition across countries and regions reflect initial differences in geographical

[161] Unlike the rise in the industrial demand for education in the UK, education was not expanded to a similar degree in India in the 19th Century. As noted by Aparna Basu (1974), during the nineteenth century the state of education in India was characterized by a relatively large university sector, aimed at producing skilled bureaucrats rather than industrialists, alongside widespread illiteracy of the masses. The literacy rate was very low, (e.g., 10% in Bengal in 1917-8) but nevertheless, attempts to expand primary education in the twentieth century were hampered by poor attendance and high drop out rates, which may suggest that the rate of return to education was relatively low. The lack of broad based education in India can also be seen using the data of Barro and Lee (2000). Despite an expansion of education throughout the twentieth century Barro and Lee report that in 1960 72.2 percent of Indians aged 15 and above had "no schooling" compared with 2 percent in the UK.

[162] Another interesting case study providing supporting evidence for the proposed hypothesis is the economic integration of the Israeli and the West Bank economies in the aftermath of the 1967 war. Trade and factor mobility between the skilled abundant economy of Israel and the unskilled abundant economy of the West Bank shifted the West Bank economy toward further specialization in the production of primary goods, and possibly triggered the astonishing increase in crude births rates from 22 per 1000 people in 1968 to 42 per 1000 in 1990, despite a decline in mortality rates. The gains from trade and development in the West Bank economy were converted primarily into an increase in population size, nearly doubling the population in those two decades. Estimates of the growth rates of output per capita over this period are less reliable and suggest that the increase was about 30%. Consistent with the proposed theory, the Palestinian uprising in the early 1990s and the gradual disintegration of the two economies resulted in the reduction in the crude birth rates.

factors and historical accidents and their manifestation in variations in institutional, demographic, and cultural factors, trade patterns, colonial status, and public policy.

References

[1] Abramowitz, M. (1993), "The Search of the Sources of Growth: Areas of Ignorance, Old and New", *Journal of Economic History* 53: 217-243.

[2] Abramovitz, M., and P.A. David, (2000), "American Macroeconomic Growth in the Era of Knowledge-Based Progress: The Long-Run Perspective", in: S.L. Engerman, and R.E. Gallman, eds., *The Cambridge Economic History of the United States*, Volume 2 (Cambridge University Press, New York).

[3] Acemoglu, D. P. Aghion and F. Zilibotti (2003), "Distance to Frontier, Selection, and Economic Growth," MIT.

[4] Acemoglu, D., S. Johnson and J. Robinson (2002), "Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution", *Quarterly Journal of Economics* 117: 1231-1294.

[5] Acemoglu, D., S. Johnson and J. Robinson (2003). "The Rise of Europe: Atlantic Trade, Institutional Change and Economic Growth", MIT.

[6] Acemoglu, D., and J.A. Robinson (2000), "Why Did the West Extend the Franchise? Democracy, Inequality and Growth in Historical Perspective", *Quarterly Journal of Economics* 115: 1167-1199.

[7] Acemoglu, D., and F. Zilibotti (1997), "Was Prometheus Unbound by Chance? Risk, Diversification, and Growth", *Journal of Political Economy* 105: 709-751.

[8] Aghion, P., and P. Howitt (1992), "A Model of Growth through Creative Destruction", *Econometrica* 60: 323-351.

[9] Alesina, A., A Devleeschauwer, W. Easterly, S. Kurlat and R. Wacziarg (2003), "Fractionalization," *Journal of Economic Growth*, 8; 155-194.

[10] Allen, R. (2000), "Economic Structure and Agricultural Productivity in Europe 1300-1800," *European Review of Economic History*, 1-25.

[11] Anderson, R. D. (1975), *Education in France 1848-1870* (Clarendon Press, Oxford).

[12] Azariadis, C. (1996), "The Economics of Poverty Traps," *Journal of Economic Growth*," 1: 449-486.

[13] Baldwin, R.E., M. Philippe and G.I.P. Ottaviano (2001), "Global Income Divergence, Trade and Industrialization: The Geography of Growth Take-Offs", *Journal of Economic Growth* 6: 5-37.

[14] Bairoch, P. (1974), "Geographical Structure and Trade Balance of European Foreign Trade From 1800-1970", *Journal of European Economic History* 3: 557-608.

[15] Bairoch, P. (1982), "International Industrialization Levels from 1750-1980", *Journal of European Economic History* 11: 269-333.

[16] Bairoch, P. (1988). *Cities and Economic Development* (The University of Chicago Press, Chicago).

[17] Barro, R.J., and G.S. Becker (1989), "Fertility Choice in a Model of Economic Growth", *Econometrica* 57: 481-501.

[18] Barro R.J., and J. Lee (2000), "International Data on Educational Attainment: Updates and Implications", Harvard University.

[19] Barro, R.J., and X. Sala-i-Martin (2003), *Economic Growth* (MIT Press, Cambridge).

[20] Basu, A. (1974), *The Growth of Education and Political Development in India 1898-1920* (Oxford University Press, Oxford).

[21] Becker, G.S. (1981), *A Treatise on the Family* (Harvard University Press, Cambridge).

[22] Becker, G.S. H.G. Lewis (1973), "On the Interaction between the Quantity and Quality of Children", *Journal of Political Economy* 81: S279-S288.

[23] Becker, G.S., K. Murphy and R. Tamura (1990),. "Human Capital, Fertility, and Economic Growth," *Journal of Political Economy,* October *98*: S12-S37.

[24] Benabou, R. (2000), "Unequal Societies: Income Distribution and the Social Contract", *American Economic Review* 90: 96-129.

[25] Ben Porath, Y. (1967), "The Production of Human Capital and the Life Cycle of Earnings," *Journal of Political Economy*, 75: 352-65.

[26] Berdugo, B., J. Sadik and N. Sussman (2003) "Delays in Technology Adoption, Appropriate Human Capital, Natural Resources and Growth" Hebrew University.

[27] Berghahn, V. R. (1994), *Imperial Germany, 1871-1914: Economy, Society, Culture and Politics* (Berghahn Books,Providence).

[28] Bertocchi, G. (2003), "The Law of Primogeniture and the Transition from Landed Aristocracy to Industrial Democracy", CEPR Discussion Paper 3723.

[29] Bertocchi, G. and F. Canova (2002), "Did Colonization Matter for Growth? An Empirical Exploration into the Historical Causes of Africa's Underdevelopment," *European Economic Review* 46: 1851-1871.

[30] Bertocchi, G. and M. Spagat, (2004), "The Evolution of Modern Educational Systems: Technical vs. General Education, Distributional Conflict, and Growth," *Journal of Development Economics*, 73: 559-582.

[31] Bisin, A., and T. Verdier (2000), "Beyond the Melting Pot: Cultural Transmission, Marriage, and the Evolution of Ethnic and Religious Traits", *Quarterly Journal of Economics* 115: 955-988.

[32] Bloom, D. E., D. Canning, and J. Sevilla (2003), "Geography and Poverty Traps," *Journal of Economic Growth* 4, 355-378.

[33] Bloom, D. and J. G. Williamson (1998), Demographic Transition and Economic Miracles in Emerging Asia, *World Bank Economic Review* 12: 419-455.

[34] Boldrin, M. and L. Jones (2002), "Mortality, Fertility, and Saving in a Malthusian Economy," *Review of Economic Dynamics* 5: 775-814.

[35] Boserup, E., (1965). *The Conditions of Agricultural Progress*, (Aldine Publishing Company, Chicago).

[36] Boucekkine, R, D. de la Croix and O. Licandro.(2003), "Early Mortality Declines at the Dawn of Modern Growth," *Scandinavian Journal of Economics*, 105: 401-418.

[37] Bourguignon, F., and T. Verdier (2000), "Oligarchy, Democracy, Inequality and Growth", *Journal of Development Economics* 62: 285-313.

[38] Bowen, J. (1981), *A History of Western Education. Vol. 3: The Modern West Europe and the New World* (St. Martin's Press, New York).

[39] Bowles, S. (1998), "Endogenous Preferences: The Cultural Consequences of Markets and other Economic Institutions," *Journal of Economic Literature* 36: 75-111.

[40] Bowles, S., and H. Gintis (1975), "Capitalism and Education in the United States", *Socialist Revolution* 5: 101-138.

[41] Boyd, R., and P.J. Richardson (1985), *Culture and the Evolutionary Process* (University of Chicago Press, Chicago).

[42] Boyer, G. (1989), "Malthus was Right After All: Poor Relief and Birth Rates in South-Eastern England", *Journal of Political Economy* 97: 93-114.

[43] Brezis, E. S , P. R Krugman, and D. Tsiddon (1993). "Leapfrogging in International Competition: A Theory of Cycles in National Technological Leadership," *American Economic Review*, 83: 1211-1219.

[44] Browning, M., L.P. Hansen and J.J. Heckman (1999),"Micro Data and General Equilibrium Models", in J. Taylor and M. Woodford (eds.),*Handbook of Macroeconomics* (North-Holland, Amsterdam).

[45] Caldwell, W. J, (1976), "Toward a Restatement of Demographic Transition Theory", *Population and Development Review* 2: 321-66.

[46] Caselli, F., (1999), "Technological Revolutions" *American Economic Review* 89: 78- 102.

[47] Caselli, F. and J. Coleman (2002), "The World Technological Frontier," Harvard University.

[48] Cavalcanti, T., and J. Tavares, (2003), "Women Prefer Larger Governments: Female Labor Supply and Public Spending," University Nova de Lisboa.

[49] Cavalli-Sforza, L.L., and M.W. Feldman (1981), *Cultural Transmission and Evolution: A Quantitative Approach* (Princeton University Press, Princeton).

[50] Cervellati, M. and U. Sunde (2003), "Human Capital Formation, Life Expectancy and the Process of Development," Iza.

[51] Chanda, A. and C.-J. Dalgaard (2003), "Dual Economies and International Total Factor Productivity Differences," University of Copenhagen

[52] Chaudhuri, K.N. (1966), "India's Foreign Trade and the Cessation of the East India Company's Trading Activities, 1828:40", *Economic History Review* 19: 345-363.

[53] Chaudhuri, K.N. (1983), "Foreign Trade and Balance of Payments (1757-1947)", in: D. Kumar, ed., *The Cambridge Economic History of India* (Cambridge University Press, Cambridge).

[54] Chesnais, J., (1992), *"The Demographic Transition: Stages, Patterns and Economic Implications"* (Clarendon Press, Oxford).

[55] Cipolla, (1969),. *Literacy and Development in the West* (Penguin Books, Harmondsworth, Middlesex)

[56] Clark, G. (1987), "Why Isn't the Whole World Developed?: Lessons from the Cotton Mills", *Journal of Economic History* 47: 141-174.

[57] Clark, G. (1991), "Labour Productivity in English Agriculture, 1300-1860", in: B.M.S. Campbell and M. Overton, eds., *Agricultural Productivity in the European Past* (Manchester University Press, Manchester).

[58] Clark (2001), "The Secret History of the Industrial Revolution." UC Davis.

[59] Clark, G. (2002), "Farmland Rental Values and Agrarian History: England and Wales, 1500-1912", *European Review of Economic History* 6: 281:308.

[60] Clark, G. (2003), "The Condition of the Working-Class in England, 1200-2000: Magna Carta to Tony Blair", UC Davis.

[61] Clark, G. and G. Hamilton (2003) "Survival of the Fittest? Capital, Human Capital, and Reproduction in European Society before the Industrial Revolution" UC Davis.

[62] Clemens, M.A., and J.G. Williamson (2004), "Why Did The Tariff-Growth Correlation Reverse After 1950?", *Journal of Economic Growth*, 9, 5-46.

[63] Coale, A.J., and R. Treadway (1986), "A Summary of the Changing Distribution of Overall Fertility, Marital Fertility, and the Proportion Married in the Provinces of Europe", in: A J. Coale and S. Watkins, eds., *The Decline of Fertility in Europe* (Princeton University Press, Princeton).

[64] Coatsworth, J. H. (1993), "Notes on the Comparative Economic History of Latin America and the United States." In Bernecker, W. and H. W. Tobler, Eds. *Development and Underdevelopment in America* (Walter de Gruyter, New York).

[65] Cody, M.L. (1966), "A General Theory of Clutch Size", *Evolution* 20: 174-184.

[66] Connolly, M. and P. F. Peretto (2003), "Industry and the Family: Two Engines of Growth," *Journal of Economic Growth* 8: 115-148.

[67] Crafts, N.F.R (1985), *British Economic Growth during the Industrial Revolution* (Oxford University Press, Oxford).

[68] Crafts, N.F.R., and C.K. Harley (1992)), "Output Growth and the Industrial Revolution: A Restatement of the Crafts-Harley View", *Economic History Review* 45: 703-730.

[69] Craig, F.W.S. (1989), *British Electoral Facts, 1832-1987* (Gower Press, Brookfield).

[70] Cressy, D. (1980), *Literacy and the Social Order: Reading and Writing in Tudor and Stuart England*, (Cambridge University Press, Cambridge).

[71] Cressy, D. (1981), "Levels of Illiteracy in England 1530-1730", in Graff H. J. ed. *Literacy and Social Development in the West: A Reader*, (Cambridge University Press, Cambridge).

[72] Cubberly, E.P. (1920), *The History of Education.*(Cambridge University Press, Cambridge).

[73] Dahan, M., and D. Tsiddon (1998), "Demographic Transition, Income Distribution, and Economic Growth", *Journal of Economic Growth* 3: 29-52.

[74] Dalgaard C-J., and C. T. Kreiner (2001), "Is Declining Productivity Inevitable?" *Journal of Economic Growth*, 6: 187-204.

[75] Darwin, C, (1859), *On the Origin of Species by Means of Natural Selection* (John Murray, London).

[76] Darwin, C. (1871), *The Descent of Man, and Selection in Relation to Sex* (John Murray, London).

[77] Dawkins, R. (1989), *The Selfish Gene* (Oxford University Press, Oxford).

[78] De la Croix D. and M. Doepke (2003), "Inequality and Growth: Why Differential Fertility Matters, "*American Economic Review*, 93: 1091-1113.

[79] Deninger K, and Squire, L (1998), "New Ways of Looking at Old Issues: Inequality and Growth", *Journal of Development Economics,* 57: 259-287.

[80] De Vries, J. (1984). *European Urbanization, 1500-1800* (Harvard University Press, Cambridge).

[81] Diamond, J. (1997), *Guns, Germs, and Steel: The Fates of Human Societies.* (Norton, New York).

[82] Doepke, M. (2004), "Accounting for Fertility Decline During the Transition to Growth", *Journal of Economic Growth* 9:

[83] Doepke, M (2005) "Child Mortality and Fertility Decline: Does the Barro-Becker Model Fit the Facts?," *Journal of Population Economics* (forthcoming).

[84] Doepke, M. and F. Zilibotti (2003) "The Macroeconomics of Child Labor Regulation," IIES, Stockholm University.

[85] Doms, M., T. Dunne and K.R. Troske (1997), "Workers, Wages and Technology", *Quarterly Journal of Economics* 112: 253-290.

[86] Duffy, J., C. Papageorgiou and F. Perez-Sebastian, (2004), Capital-Skill Complementarity? Evidence from a Panel of Countries, *Review of Economic and Statistics.*

[87] Durham, W. (1982), "Interaction of Genetic and Cultural Evolution: Models and Examples", *Human Ecology* 10: 289-323.

[88] Durlauf, S. N. (1996), "A Theory of Persistent Income Inequality," *Journal of Economic Growth*, 1: 75-94.

[89] Durlauf, S.N. and P. A. Johnson (1995), "Multiple Regimes and Cross-Country Growth Behavior," *Journal of Applied Econometrics* 10: 365-84."

[90] Durlauf, S.N. and D. Quah (1999), "The New Empirics of Economic Growth" in *Handbook of Macroeconomics*, J. B. Taylor and M. Woodford (eds.), (North-Holland, Amsterdam).

[91] Dyson, T., and M. Murphy (1985), "The Onset of Fertility Transition", *Population and Development Review* 11: 399-440.

[92] Eckstein, Z., P. Mira and K. I. Wolpin.(1999), "A Quantitative Analysis of Swedish Fertility Dynamics:1751-1990." *Review of Economic Dynamics* 2: 137-165.

[93] Easterlin, R. (1981), "Why Isn't the Whole World Developed?", *Journal of Economic History* 41: 1-19.

[94] Easterly, W. and R. Levine (1997), "Africa's Growth Tragedy: Policies and Ethnic Divisions," *Quarterly Journal of Economics* 111: 1203-1250.

[95] Easterly, W. and R. Levine (2003), "Tropics, germs, and crops: the role of endowments in economic development", *Journal of Monetary Economics*, 50: 3-39.

[96] Edlund, L. and N-P. Lagerlof (2002), "Implications of Marriage Institutions for Redistribution and Growth," Columbia University.

[97] Endler, J.A. (1986), *Natural Selection in the Wild* (Princeton University Press, Princeton).

[98] Engerman, S., and K.L. Sokoloff (2000), "Factor Endowment, Inequality, and Paths of Development Among New World Economies", UCLA.

[99] Erlich, I., and F. T. Lui (1991), "Intergenerational trade, Longevity, and Economic Growth," *Journal of Political Economy 99: 1029-1059.*

[100] Estavadeordal, A., B. Frantz and A.M. Taylor (2002), "The Rise and Fall of World Trade, 1870-1939", *Quarterly Journal of Economics*118: 359-407.

[101] Estavadeordal, A., and A.M. Taylor (2002), "A Century of Missing Trade", *American Economic Review* 92: 383-393.

[102] Estevo, G. (1983), *The Struggle for Rural Mexico.* (Bergin and Garvey, South Hadley, MA)..

[103] Feinstein, C.H. (1972), *National Income, Expenditure and Output of the United Kingdom 1855-1965*, (Cambridge University Press, Cambridge).

[104] Fernandez, R., A. Fogli and C. Olivetti (2004), "Mothers and Sons: Preference Formation and Female Labor Force Dynamics," *Quarterly Journal of Economics* 119.

[105] Fernandez, R., and R. Rogerson (1996): "Income Distribution, Communities, and the Quality of Public Education," *Quarterly Journal of Economics*, 111: 135-164.

[106] Fernandez-Villaverde J. (2003), "Was Malthus Right? Economic Growth and Population Dynamics", University of Pennsylvania.

[107] Feyrer, J. (2003), "Convergence by Parts," Dartmouth College.

[108] Fiaschi D. and A. M. Lavezzi (2003), "Distribution Dynamics and Nonlinear Growth," *Journal of Economic Growth* 4: 379-402.

[109] Field, A. (1976), "Educational Reform and Manufacturing Development in Mid-Nineteenth Century Massachusetts", *Journal of Economic History* 36: 263-266.

[110] Findlay, R., and H. Keirzkowsky (1983), "International Trade and Human Capital: A simple General Equilibrium Model", *Journal of Political Economy* 91: 957-978

[111] Findlay, R., and K.H. O'Rourke (2003), "Commodity Market Integration, 1500-2000", in: M.D. Bordo, A.M. Taylor and J.G. Williamson, eds., *Globalization in Historical Perspective* (The University of Chicago Press, Chicago).

[112] Flora, P., F. Kraus and W. Pfenning (1983), *State, Economy and Society in Western Europe 1815-1975, Vol. I.* (St. James Press, Chicago).

[113] Fogel, R. W.(1994), "Economic Growth, Population Theory, and Physiology: The Bearing of Long-Term Processes on the Making of Economic Policy," *American Economic Review* 84: 369-95.

[114] Foster, A.D., and M.R. Rosenzweig (1996), "Technical Change and Human-Capital Returns and Investments: Evidence from the Green Revolution", *American Economic Review* 86: 931-953.

[115] Galor, O. (1996), "Convergence?: Inferences from Theoretical Models," *Economic Journal*, 106: 1056-1069.

[116] Galor, O., and O. Moav (2000), "Ability Biased Technological Transition, Wage Inequality and Growth", *Quarterly Journal of Economics* 115: 469-498.

[117] Galor, O., and O. Moav (2002), "Natural Selection and the Origin of Economic Growth", *Quarterly Journal of Economics* 117: 1133-1192.

[118] Galor, O., and O. Moav (2003), "Das Human Kapital: A Theory of the Demise of the Class Structure", Brown University.

[119] Galor, O., and O. Moav (2004a), "From Physical to Human Capital Accumulation: Inequality and the Process of Development", *Review of Economic Studies,*

[120] Galor, O., and O. Moav (2004b), "Natural Selection and the Evolution of Life Expectancy," Hebrew University.

[121] Galor, O., O. Moav and D. Vollrath (2003), "Land Inequality and the Origin of Divergence and Overtaking in the Growth Process: Theory and Evidence," Brown University.

[122] Galor, O., and A. Mountford, (2003), "Trading Population for Productivity," Brown University.

[123] Galor, O., and D. Tsiddon (1997), "Technological Progress, Mobility, and Growth", *American Economic Review* 87: 363-382.

[124] Galor, O., and D.N. Weil (1996), "The Gender Gap, Fertility, and Growth", *American Economic Review* 86: 374-387.

[125] Galor, O., and D.N. Weil (1999), "From Malthusian Stagnation to Modern Growth", *American Economic Review*, 89: 150-154.

[126] Galor, O., and D.N. Weil (2000), "Population, Technology and Growth: From the Malthusian regime to the Demographic Transition", *American Economic Review* 110: 806-828.

[127] Galor, O., and J. Zeira (1993), "Income Distribution and Macroeconomics", *Review of Economic Studies* 60: 35-52.

[128] Gallup, J. L., J. D. Sachs, and A. D. Mellinger, "Geography and Economic Development," NBER Working Paper No. w6849, December 1998.

[129] Goldin, C. (1990), *Understanding The Gender Gap: An Economic History of American Women.* (Oxford University Press, New York)

[130] Goldin, C. (2001), "The Human Capital Century and American Leadership: Virtues of the Past", *Journal of Economic History* 61: 263-292.

[131] Goldin, C., and L.F. Katz (1998), "The Origins of Technology-Skill Complementarity", *Quarterly Journal of Economics* 113: 693-732.

[132] Goldin, C., and L.F. Katz (2001), "On the Legacy of U.S. Educational Leadership: Notes on Distribution and Economic Growth in the 20th Century", *American Economic Review* 91: 18-23.

[133] Goodfriend, M., and J. McDermott (1995), "Early Development", *American Economic Review* 85: 116-133.

[134] Gould, E.D., O. Moav and A. Simhon (2003), "The Mystery of Monogamy," Hebrew University.

[135] Gould, S.J. (1977), *Ever Since Darwin* (Norton, New York).

[136] Grant, B.R., and P.R. Grant (1989), *Evolutionary Dynamics of a Natural Population* (University of Chicago Press, Chicago).

[137] Green, A. (1990), *Education and State Formation* (St. Martin's Press, New York).

[138] Greenwood, J., and A. Seshadri (2002), "The U.S. Demographic Transition," *American Economic Review* 92: 153-159.

[139] Grossman, G.M., and E. Helpman (1991), *Innovation and Growth* (MIT Press, Cambridge).

[140] Grossman, H. I., and M. Kim (1999), "Education Policy: Egalitarian or Elitist?" *Economics and Politics* 15: 225-246.

[141] Gylfason T. (2001)."Natural Resources, Education, and Economic Development," *European Economic Review* 45: 847-859.

[142] Hansen, G., and E. Prescott (2002), "Malthus to Solow", *American Economic Review* 92: 1205-1217.

[143] Hansson, I., and C. Stuart (1990), "Malthusian Selection of Preferences", *American Economic Review* 80: 529-544.

[144] Hanushek, E.A. (1992), "The Trade-Off between Child Quantity and Quality", *Journal of Political Economy* 100: 84-117.

[145] Hassler, J., and J.V. Rodriguez Mora (2000), "Intelligence, Social Mobility, and Growth", *American Economic Review* 90: 888-908.

[146] Hazan, M., and B. Berdugo (2002), "Child Labor, Fertility and Economic Growth," *Economic Journal*, 112: 810-828.

[147] Hazan, M., and H. Zoabi (2004). "Does Longevity Cause Growth," Hebrew University.

[148] Hechter, M. (2001), Internal Colonialism Study: National Integration in the British Isles, 1851-1966 (Inter-University Consortium for Political and Social Research, Ann Arbor).

[149] Hernandez, D. J. (2000), *Trends in the Well Being of America's Children and Youth*. U.S. Bureau of the Census.

[150] U.S. Bureau of the Census, (1975), *Historical Statistics of the United States: Colonial Times to 1970, Part 1*, (Washington D.C.).

[151] Hibbs D. A., and O. Olson (2004), "Biogeography and Long-Run Economic Development", *European Economic Review* 48:

[152] Horrell, S., and J. Humphries (1995), "The Exploitation of Little Children": Child Labor and the Family Economy in the Industrial Revolution," *Exploration in Economic History* 32: 485-516

[153] Howitt, P., and D. Mayer-Foulkes (2002), "R&D, Implementation and Stagnation: A Schumpeterian Theory of Convergence Clubs", Brown University.

[154] Hurt, J. (1971), *Education in Evolution* (Paladin, London).

[155] Iyigun, M. F. (2004), "Geography, Demography, and Early Development," Boulder.

[156] Jones C.I. (1997), "Convergence Revisited," *Journal of Economic Growth*, 2: 131-154.

[157] Jones C.I. (2001), "Was an Industrial Revolution Inevitable? Economic Growth Over the Very Long Run" *Advances in Macroeconomics 1*: 1-43.

[158] Jones, E.L. (1981), *The European Miracle: Environments, Economies and Geopolitics in the History of Europe and Asia* (Cambridge University Press, Cambridge).

[159] Kalemli-Ozcan, S. (2002), "Does the Mortality Decline Promote Economic Growth," *Journal of Economic Growth* 7: 411-439.

[160] Kalemli-Ozcan, S., H.E. Ryder and D. N. Weil (2000), "Mortality Decline,Human Capital Investment,and Economic Growth" *Journal of Development Economics*, 62: 1-23

[161] Kaplan, H.S. (1994), "Evolutionary and Wealth Flows of Fertility - Empirical Tests and new Models", *Population and Development Review* 20: 753-791.

[162] Kettlewell, H.B.D. (1973), *The Evolution of Melanism* (Clarendon Press, Oxford).

[163] Knick, H.C. (1999), "Reassessing the Industrial Revolution: a Macro View ", in: J. Mokyr, ed., *The British Industrial Revolution: an Economic Perspective* (Westview Press, Boulder).

[164] Kogel, T., and A. Prskawetz. (2001), "Agricultural Productivity Growth and Escape from Malthusian Trap", *Journal of Economic Growth* 6: 337-357.

[165] Kohler, H., J.L. Rodgers and K. Christensen (1999), "Is Fertility Behavior in our Genes? Findings from a Danish Twin Study", *Population and Development Review* 25: 253-263.

[166] Komlos, J. and M. Artzrouni (1990), "Mathematical Investigations of the Escape from the Malthusian Trap", *Mathematical Population Studies* 2: 269-287.

[167] Kremer, M. (1993), "Population Growth and Technological Change: One Million B.C. to 1990," *Quarterly Journal of Economics* 108: 681-716.

[168] Kremer, M. and D. L. Chen (2002), "Income Distribution Dynamics with Endogenous Fertility," *Journal of Economic Growth* 7: 227-258

[169] Krugman, P., and A. Venables (1995), "Globalization and the inequality of nations", *Quarterly Journal of Economics* 90: 857-880.

[170] Kuhn, T.S., (1957), *The Copernican Revolution,* (Cambridge, MA)

[171] Kurian, G.T. (1994), *Datapedia of the U.S. 1790-2000, America Year by Year* (Bernan Press, Lonham).

[172] Kuznets, S. (1967), "Quantitative Aspects of the Economic Growth of Nations : X. Level and Structure of Foreign Trade: Long-Term Trends", *Economic Development and Cultural Change* 15: 1-140.

[173] Kuzynski, R. R. (1969), *The Measurement of Population Growth*, (Gordon and Breach Science Publishers, New York).

[174] Lack, D. (1954), *The Natural Regulation of Animal Numbers* (Clarendon Press, Oxford).

[175] Lagerlof, N., (2003a), "From Malthus to Modern Growth: The Three Regimes Revisited", *International Economic Review* 44: 755-777.

[176] Lagerlof, N. (2003b), "Gender Equality and Long-Run Growth," *Journal of Economic Growth*, 8: 403-426.

[177] Landes, D.S. (1969), *The Unbound Prometheus. Technological Change and Industrial Development in Western Europe from 1750 to the Present* (Cambridge University Press, Cambridge).

[178] Landes, D. S. (1998), *The Wealth and Poverty of Nations* (Norton, New York).

[179] Lee, C. (1979), *British Regional Employment Statistics,1841-1971* (Cambridge University Press, Cambridge).

[180] Lee, R.D. (1997), "Population Dynamics: Equilibrium, Disequilibrium, and Consequences of Fluctuations", in: O. Stark and M. Rosenzweig, eds., *The Handbook of Population and Family Economics* (Elsevier, Amsterdam).

[181] Levy-Leboyer, M., and F. Bourguignon (1990), *The French Economy in the Nineteenth Century* (Cambridge University Press, Cambridge).

[182] Lewis, W.A. (1954), "Economic Development with Unlimited Supply of Labor", *The Manchester School* 22: 139-191.

[183] Lindert, P.H., and J.G. Williamson (1976), "Three Centuries of American Inequality", *Research in Economic History* 1: 69-123.

[184] Livi-Bacci, M. (1997), *A Concise History of World Population* (Blackwel, Oxford).

[185] Livingston F, (1958), "Anthropological Implications of Sickle Cell Distribution in West Africa", *American Anthropologist* 60: 533-562.

[186] Lucas, R.E. (2002), *The Industrial Revolution: Past and Future* (Harvard University Press, Cambridge).

[187] MacArthur, R.H., and E.O. Wilson (1967), *The Theory of Island Biogeography* (Princeton University Press, Princeton).

[188] Maddison, A. (2001), *The World Economy: A Millennia Perspective* (OECD, Paris).

[189] Maddison, A. (2003), *The World Economy: Historical Statistics* CD-ROM (OECD, Paris).

[190] Malthus, T.R. (1798), *An Essay on the Principle of Population* (printed for J. Johnson, in St. Paul's Church-Yard, London).

[191] Masters, W.E., and M.S. McMillan (2001), "Climate and Scale in Economic Growth," *Journal of Economic Growth* 6: 167-187.

[192] Matsuyama, K. (1992), "Agricultural Productivity, Comparative Advantage, and Economic Growth", *Journal of Economic Theory* 58: 317-334.

[193] Matthews, R.C., C.H. Feinstein and J.C. Odling-Smee (1982), *British Economic Growth 1856-1973* (Stanford University Press, Stanford).

[194] McClelland, C.E. (1980), *State, Society, and University in Germany: 1700-1914* (Cambridge University Press, Cambridge).

[195] McCloskey, D.N. (1981), "The Industrial Revolution, A Survey", in: R.C. Floud and D.N. McCloskey, eds., *The Economic History of Britain Since 1700 Vol 1* (Cambridge University Press, Cambridge).

[196] McDermott, J. (2002), "Development Dynamics: Economic Integration and the Demographic Transition", *Journal of Economic Growth* 7:371-410.

[197] McEvedy, C., and R. Jones (1978), *Atlas of World Population History* (Penguin, London).

[198] Menchik, P., and M. David (1983), "Income Distribution, Lifetime Savings, and Bequests", *American Economic Review* 73: 672–690.

[199] Mitch, D. (1992), *The Rise of Popular Literacy in Victorian England: The Influence of Private Choice and Public Policy* (University of Pennsylvania Press, Philadelphia).

[200] Mitch, D. (1993), "The Role of Human Capital in the First Industrial Revolution", in: J. Mokyr, ed., *The British Industrial Revolution: An Economic Perspective* (Westview Press, Boulder).

[201] Mitch, D. (2001), "The Rise of Mass Education and Its contribution to economic Growth in Europe, 1800-2000", mimeo (University of Maryland Baltimore County).

[202] Mitchell, B. (1981), *European Historical Statistics*, 2nd ed., (New York).

[203] Moav, O. (2005), "Cheap Children and the Persistence of Poverty", *Economic Journal*, 114:

[204] Mokyr, J. (1985), *Why Ireland Starved : A Quantitative and Analytical History of the Irish Economy, 1800-1850* (Allen and Unwin, London).

[205] Mokyr, J. (1990), *The Lever of Riches* (Oxford University Press, New York).

[206] Mokyr, J. (1993), "The New Economic History and the Industrial Revolution," in: J. Mokyr, ed., *The British Industrial Revolution: an Economic Perspective* (Westview Press, Boulder).

[207] Mokyr, J. (2001), "The Rise and Fall of the Factory System: Technology, Firms, and Households since the Industrial Revolution", *Carnegie-Rochester Conference Series On Public Policy* 55: 1-45.

[208] Mokyr, J. (2002), *The Gifts of Athena: Historical Origins of the Knowledge Economy* (Princeton University Press, Princeton).

[209] Mookherjee, D., and D. Ray (2003), "Persistent Inequality", *Review of Economic Studies* 70: 369–393.

[210] Morrisson, C., and W. Snyder (2000), "Income Inequalities in France from the Early Eighteenth Century to 1985", *Revue Economique* 51: 119-154.

[211] Mountford, A., (1998) "Trade, Convergence and Overtaking," *Journal of International Economics*, 46: 167-182.

[212] Muller, D.K. (1987), "The Process of Systematization: the Case of German Secondary Education", in: D. Muller, F. Ringer and B. Simon, eds., *The Rise of the Modern Educational System* (Cambridge University Press, Cambridge).

[213] Neher, A.P. (1971), "Peasants, Procreation and Pensions," *American Economic Review* 61: 380-389.

[214] Nelson, R.R., and E.S. Phelps (1966), "Investment in Humans, Technological Diffusion, and Economic Growth", *American Economic Review* 56: 69-75.

[215] North, D. C. (1981), *Structure and Change in Economic History*, (W.W. Norton & Co., New York).

[216] O'Rourke, K.H., A.M. Taylor and J.G. Williamson (1996), "Factor Price Convergence in the Late Nineteenth Century", *International Economic Review* 37: 499-530.

[217] O'Rourke, K.H., and J.G. Williamson (1999), *Globalization and History* (MIT Press, Cambridge).

[218] O'Rourke, K.H., and J.G. Williamson (2003), "Malthus to Ohlin," Harvard University.

[219] Parente, S., and E.C. Prescott (2000), *Barriers to Riches* (MIT Press, Cambridge).

[220] Pereira, A.S. (2003), "When Did Modern Economic Growth Really Start? The Empirics of Malthus to Solow." UBC.

[221] Persson, T., and G. Tabellini (1994), "Is Inequality Harmful for Growth?", *American Economic Review* 84: 600-621.

[222] Pollard, S. (1963), "Factory Discipline in the Industrial Revolution", *Economic History Review* 16: 254-271.

[223] Pomeranz, K. (2000), *The Great Divergence: China, Europe and the Making of the Modern World Economy* (Princeton University Press, Princeton).

[224] Pritchett, L.(1997),."Divergence, Big Time,". *Journal of Economic Perspectives* 11: 3-17.

[225] Przeworski, A. (2003), "The Last Instance:Are Institutions the Primary Cause of Economic Development?" Department of Politics, New York University.

[226] Psacharopoulos, G. and H. A. Patrinos (2002), "Returns to Investment in Education: A Further Update September 1, 2002," World Bank.

[227] Quah, D. (1996), "Convergence Empirics Across Economies with (some) Capital Mobility," *Journal of Economic Growth*, 1: 95-124.

[228] Quah, D. (1997), "Empirics for Growth and Distribution: Stratification, Polarization, and Convergence Clubs," *Journal of Economic Growth*, 2: 27-61.

[229] Razin, A., and U. Ben-Zion (1975), "An Intergenerational Model of Population Growth," *American Economic Review*, 65: 923-933.

[230] Ringer, F. (1979), *Education and Society in Modern Europe* (Indiana University Press, Bloomington).

[231] Robson, A.J. (2001), "The Biological Basis of Economic Behavior", *Journal of Economic Literature* 39: 11-33.

[232] Robson A.J., and H. S. Kaplan (2003), "The Evolution of Human Longevity and Intelligence in Hunter-Gatherer Economies," *American Economic Review* 93: 150-169.

[233] Rodgers, J.L., and D. Doughty (2000), "Genetic and Environmental Influences on Fertility Expectations and Outcomes Using NLSY Kinship Data", in: J.L. Rodgers, D.C. Rowe and W.B. Miller, eds., *Genetic Influences on Human Fertility and Sexuality* (Kluwer, Boston).

[234] Rodgers, J.L., K. Hughes, H. Kohler, K. Christensen, D. Doughty, D.C. Rowe and W.B. Miller (2001a), "Genetic influence helps explain variation in human fertility: Evidence from recent behavioral and molecular genetic studies", *Current Directions in Psychological Science* 10: 184-188.

[235] Rodgers, J.L., H. Kohler, K. Ohm Kyvik, and K. Christensen (2001b), "Behavior Genetic Modeling of Human Fertility: Findings from a Contemporary Danish Twin Study", *Demography* 38: 29-42.

[236] Rodriguez R. and D. Rodrik (2001), "Trade Policy and Economic Growth: A Skeptic's Guide to the Cross-National Evidence," NBER *Macroeconomics Annual* eds. B. Bernanke and K. S. Rogoff, (MIT Press, Cambridge).

[237] Rodrik D., A. Subramanian and F. Trebbi (2004), "Institutions Rule: The Primacy of Institutions over Geography and Integration in Economic Development, *Journal of Economic Growth* 9.

[238] Rosenzweig, M.R., and K. I. Wolpin (1980), "Testing the Quantity-Quality Fertility Model: the use of Twins as a Natural Experiment", *Econometrica* 48: 227-240.

[239] Saint-Paul, G. (2003), *"*On Market and Human Evolution*"* CEPR Discussion Paper No. 3654.

[240] Sala-i-Martin, X. (2002), The Disturbing "Rise" in World Income Distribution," NBER Working Paper.

[241] Sanderson, M. (1995), *Education, Economic Change and Society in England 1780-1870* (Cambridge University Press, Cambridge).

[242] Schofield, R. S. (1973) "Dimensions of Illiteracy, 1750-1850," *Explorations in Economic History* 10: 437-454.

[243] Schultz, T.W. (1964), *Transforming Traditional Agriculture* (Yale University Press, New Haven).

[244] Schultz, T.W. (1975), "The Value of the Ability to Deal with Disequilibria", *Journal of Economic Literature* 8: 827-846.

[245] Simon, B. (1987), "Systematization and Segmentation in Education: the Case of England", in: D. Muller, F. Ringer and B. Simon, eds., *The Rise of the Modern Educational System* (Cambridge University Press, Cambridge).

[246] Smith, A. (1776), *The Wealth of Nations* (Modern Library, New-York, 1937).

[247] Smith, D. (2001), "International evidence on how income inequality and credit market imperfections affect private saving rates", *Journal of Development Economics* 64: 103-127.

[248] Soares, R. R. (2003), "Mortality Reductions, Educational Attainment, and Fertility Choice," University of Maryland.

[249] Spree, R. (1977), *Die Wachstumszyklen der deutschen Wirtschaft von 1840 bis 1880* (Dunker & Humboult, Berlin).

[250] Stokey, N. (2001), "A Quantitative Model of the British Industrial Revolution, 1780-1850. *Carnegie-Rochester Conference Series on Public Policy* 55: 55-109.

[251] Stone, L., (1969), "Literacy and Education in England 1640-1900," *Past and Present* 42: 69-139.

[252] Tamura, R.F. (2002), "Human Capital and the Switch From Agriculture to Industry", *Journal of Economic Dynamics and Control* 27: 207-242.

[253] Taylor, A. M. (1999), "Sources of Convergence in the Late Nineteenth Century", *European Economic Review* 9: 1621-45.

[254] U.S. Bureau of the Census, (1975), *Historical Statistics of the United States: Colonial Times to 1970 (part 1)* Series D 830-844, p.172.

[255] Voth, H-J., (2003) "Living Standards During the Industrial Revolution: An Economist's Guide," *American Economic Review*, 93: 221-226.

[256] Voth, H-J., (2004) "Living Standards and the Urban Environment," in Paul Johnson and Roderick Floud,eds.,*The Cambridge Economic History of England*. (Cambridge University Press, Cambridge).

[257] Weil, D.N. (2004), *Economic Growth*, (Addison Weseley, Boston).

[258] Weisdorf, J. L. (2004) "From Stagnation to Growth: Revisiting Three Historical Regimes." *Journal of Population Economics* (forthcoming).

[259] Weisenfeld, S.L. (1975), "Sickle Cell Trait in Human Biological and Cultural Evolution", *Science* 157: 1135-1140.

[260] West, E.G. (1985), "Literacy and the Industrial Revolution", in: J. Mokyr, ed., *The Economics of the Industrial Revolution* (Rowman and Littlefield).

[261] Williamson, J.G. (1985), *Did British Capitalism Breed Inequality?* (Allen & Unwin, Boston).

[262] Wilson, E.O. (1975), *Sociobiology* (The Belknap Press of Harvard University Press, Cambridge).

[263] Wolthuis, J. (1999), Lower Technical Education in the Netherlands 1798-1993: the Rise and Fall of a Subsystem. Netherlands.

[264] Wrigley, E.A. (1969), *Population and History* (McGraw-Hill, New York).

[265] Wrigley, E.A., and R.S. Schofield (1981), *The Population History of England 1541-1871: A Reconstruction* (Harvard University Press, Cambridge).

[266] Yates, P.L. (1959), *Forty Years of Foreign Trade: A Statistical Handbook With Special Reference to Primary Products and Underdeveloped Countries* (Allen & Unwin, London).

[267] Young, A. (1991), "Learning by Doing and the Dynamic Effects of International Trade," *Quarterly Journal of Economics*, 106 369-405.

Poverty Traps ⋆

Prepared for the Handbook of Economic Growth
(Philippe Aghion and Steven Durlauf, eds.)

Costas Azariadis [a] and John Stachurski [b]

[a] *Department of Economics, University of California Los Angeles*
405 Hilgard Avenue, Los Angeles, CA 90095-1477

[b] *Department of Economics, The University of Melbourne, VIC 3010, Australia*

> In the problem of economic development, a phrase that crops up frequently is 'the vicious circle of poverty.' It is generally treated as something obvious, too obvious to be worth examining. I hope I may be forgiven if I begin by taking a look at this obvious concept. (R. Nurkse, 1953)

1 Introduction

Despite the considerable amount of research devoted to economic growth and development, economists have not yet discovered how to make poor countries rich. As a result, poverty remains the common experience of billions. One half of the world's people live on less than $2 per day. One fifth live on less than $1. [1] If modern production technologies are essentially free for the taking, then why is it that so many people are still poor?

⋆ This chapter draws on material contained in two earlier surveys by the first author (Azariadis 1996, 2004). Support from the Program of Dynamic Economics at UCLA is acknowledged with thanks, as is research assistance from Athanasios Bolmatis, and discussions with David de la Croix, Chris Edmond, Cleo Fleming, Oded Galor, Karla Hoff, Kirdan Lees and Yasusada Murata. The second author thanks the Center for Operations Research and Econometrics at Université Catholique de Louvain for their hospitality during a period when part of this survey was written. All simulations and estimations use the open source programming language R.

Email addresses: azariadi@ucla.edu (Costas Azariadis),
j.stachurski@econ.unimelb.edu.au (John Stachurski).

[1] Figures are based on Chen and Ravallion (2001). Using national surveys they calculate a total head-count for the $1 and $2 poverty lines of 1.175 and 2.811

The literature that we survey here contains the beginnings of an answer to this question. First, it is true that technology is the primary determinant of a country's income. However, the most productive techniques will not always be adopted: There are self-reinforcing mechanisms, or "traps," that act as barriers to adoption. Traps arise both from market failure and also from "institution failure;" that is, from traps within the set of institutions that govern economic interaction. Institutions—in which we include the state, legal systems, social norms, conventions and so on—are determined endogenously within the system, and may be the direct cause of poverty traps; or they may interact with market failure, leading to the perpetuation of an inefficient status quo.

There is no consensus on the view that we put forward. Some economists regard institutions such as the state or policy as largely exogenous. Many argue that the primary suspect for the unfortunate growth record of the least developed countries should be bad domestic policy. Bad policy can be changed directly, because it is exogenous, rather than determined within the system. Sound governance and free market forces are held to be not only necessary but also *sufficient* to revive the poor economies, and to catalyze their convergence. Because good policy is available to all, there are no poverty traps.

The idea that good policy and the invisible hand are sufficient for growth is at least vacuously true, in the sense that an all-seeing and benevolent social planner who completes the set of markets can succeed where developing country governments have failed. But this is not a theory of development, and of course benevolent social planners are not what the proponents of good governance and liberalization have in mind. Rather, their argument is that development can be achieved by the poor countries if only governments allow the market mechanism to function effectively—to get the prices right—and permit economic agents to fully exploit the available gains from trade. This requires not just openness and non-distortionary public finance, but also the enforcement of property rights and the restraint of predation. [2]

In essence, this is the same story that the competitive neoclassical benchmark economy tells us: Markets are complete, entry and exit is free, transaction costs are negligible, and technology is convex at an efficient scale relative to

billion respectively in 1998. Their units are 1993 purchasing power adjusted US dollars.

[2] Development theory then reduces to Adam Smith's famous and compelling dictum, that "Little else is requisite to carry a state to the highest degree of opulence from the lowest barbarism but peace, easy taxes, and a tolerable administration of justice."

the size of the market. As a result, the private and social returns to production and investment are equal. A complete set of "virtual prices" ensures that all projects with positive net social benefit are undertaken. Diminishing returns to the set of reproducible factor inputs implies that when capital is scarce the returns to investment will be high. The dynamic implications of this benchmark were summarized by Solow (1956), Cass (1965), and Koopmans (1965). Even for countries with different endowments, the main conclusion is convergence.

There are good reasons to expect this benchmark will have relevance in practice. The profit motive is a powerful force. Inefficient practices and incorrect beliefs will be punished by lost income. Further, at least one impetus shaping the institutional environment in which the market functions is the desire to mitigate or correct perceived social problems; and one of the most fundamental of all social problems is scarcity. Over time institutions have often adapted so as to relieve scarcity by addressing sources of market inefficiency. [3]

In any case, the intuition gained from studying the neoclassical model has been highly influential in the formulation of development policy. A good example is the structural adjustment programs implemented by the International Monetary Fund. The key components of the Enhanced Structural Adjustment Facility—the centerpiece of the IMF's strategy to aid poor countries and promote long run growth from 1987 to 1999—were prudent macroeconomic policies and the liberalization of markets. Growth, it was hoped, would follow automatically.

Yet the evidence on whether or not non-distortionary policies and diminishing returns to capital will soon carry the poor to opulence is mixed. Even relatively well governed countries have experienced little or no growth. For example, Mali rates as "free" in recent rankings by Freedom House. Although not untroubled by corruption, it scores well in measures of governance relative to real resources (Radlet 2004; Sachs et al. 2004). Yet Mali is still desperately poor. According to a 2001 UNDP report, 70% of the population lives on less than $1 per day. The infant mortality rate is 230 per 1000 births, and household final consumption expenditure is down 5% from 1980.

Mali is not an isolated case. In fact for all of Africa Sachs et al. (2004) argue that

> With highly visible examples of profoundly poor governance, for example in Zim-

[3] See Greif, Milgrom and Weingast (1994) for one of many possible examples.

babwe, and widespread war and violence, as in Angola, Democratic Republic of Congo, Liberia, Sierra Leone and Sudan, the impression of a continent-wide governance crisis is understandable. Yet it is wrong. Many parts of Africa are well governed, and yet remain mired in poverty. Governance is a problem, but Africa's development challenges are much deeper.

There is a further problem. While the sufficiency of good policy and good governance for growth is still being debated, what can be said with certainty is that they are both elusive. The institutions that determine governance and other aspects of market interaction are difficult to reform. Almost everyone agrees that corruption is bad for growth, and yet corruption remains pervasive. Some institutions important to traditional societies have lingered, inhibiting the transition to new techniques of production. The resistance of norms and institutions to change is one reason why the outcome of liberalization and governance focused adjustment lending by the IMF has often been disappointing.

We believe that in practice there are serious problems with direct application of the benchmark story. First, for reasons outlined below, numerous deviations from the neoclassical benchmark generate market failure. Because of these failures, good technologies are not always adopted, and productive investments are not always undertaken. Inefficient equilibria exist. Second, as Hoff (2000) has emphasized, the institutional framework in which market interaction takes place is not implemented "from above." Rather it is determined within the system. Bounded rationality, imperfect information, and costly transactions make institutions and other "rules of the game" critical to economic performance; and the equilibria for institutions may be inefficient.

Moreover, as we shall see, inefficient equilibria have a bad habit of *reinforcing* themselves. Corrupt institutions can generate incentives which reward more corruption. Workers with imperfectly observed skills in an unskilled population may be treated as low skilled by firms, and hence have little incentive to invest large sums in education. Low demand discourages investment in increasing returns technology, which reduces productivity and reinforces low demand. That these inefficient outcomes are self-reinforcing is important—were they not then presumably agents would soon make their way to a better equilibrium.

Potential departures from the competitive neoclassical benchmark which cause market failure are easy to imagine. One is increasing returns to scale, both internal and external. Increasing returns matter because development is almost synonymous with industrialization, and with the adoption of modern pro-

duction techniques in agriculture, manufacturing and services. These modern techniques involve both fixed costs—internal economies—and greater specialization of the production process, the latter to facilitate application of machines.

The presence of fixed costs for a *given* technology is more troubling for the neoclassical benchmark in poor countries because there market scale is relatively small. If markets are small, then the neoclassical assumption that technologies are convex at an efficient scale may be violated. The same point is true for market scale and specialization, in the sense that for poor countries a given increase in market scale may lead to considerably more opportunity to employ indirect production. [4]

Another source of increasing returns follows from the fact that modern production techniques are knowledge-intensive. As Romer (1990) has emphasized, the creation of knowledge is associated with increasing returns for several reasons. First, knowledge is non-rival and only partially excludable. Romer's key insight is that in the presence of productive non-rival inputs, *the entire replication-based logical argument for constant returns to scale immediately breaks down.* Thus, knowledge creation leads to positive technical externalities and increasing returns. Second, new knowledge tends in the aggregate to complement existing knowledge.

If scale economies, positive spillovers and other forms of increasing returns are important, then long run outcomes may not coincide with the predictions of the neoclassical benchmark. The essence of the problem is that when returns are increasing a rise in output *lowers* unit cost, either for the firm itself or for other firms in the industry. This sets in motion a chain of positive self-reinforcement. Lower unit cost encourages production, which further lowers unit cost, and so on. Such positive feedbacks can strongly reinforce either poverty *or* development.

[4] Domestic markets *are* small in many developing countries, despite the possibility of international trade. In tropical countries, for example, roads are difficult to build and expensive to maintain. In Sub-Saharan Africa, overland trade with European and other markets is cut off by the Sahara. At the same time, most Sub-Saharan Africans live in the continent's interior highlands, rather than near the coast. To compound matters, very few rivers from the interior of this part of the continent are ocean-navigable, in contrast to the geography of North America, say, or Europe (Limao and Venables 2001; Sachs et al. 2004). The potential for international trade to mitigate small market size is thus far lower than for a country with easy ocean access, such as Singapore or the UK.

Another deviation from the competitive neoclassical benchmark that we discuss at length is failure in credit and insurance markets. Markets for loans and insurance suffer more acutely than most from imperfections associated with a lack of complete and symmetric information, and with all the problems inherent in anonymous trading over time. Borrowers may default or try not to pay back loans. The insured may become lax in protecting their own possessions.

One result of these difficulties is that lenders usually require collateral from their borrowers. Collateral is one thing that the poor always lack. As a result, the poor are credit constrained. This can lead to an inefficient outcome which is self-reinforcing: Collateral is needed to borrow funds. Funds are needed to take advantage of economic opportunities—particularly those involving fixed costs. The ability to take advantage of opportunities determines income; and through income is determined the individual's wealth, and hence their ability to provide collateral. Thus the poor lack access to credit markets, which is in turn the cause of their own poverty.

An important aspect of this story for us is that many modern sector occupations and production techniques have indivisibilities which are not present in subsistence farming, handicraft production or other traditional sector activities. Examples include projects requiring fixed costs, or those needing large investments in human capital such as education and training. The common thread is that through credit constraints the uptake of new technologies is inhibited.

With regards to insurance, it has been noted that—combined with limited access to credit—a lack of insurance is more problematic for the poor than the rich, because the poor cannot self-insure by using their own wealth. As a result, a poor person wishing to have a smooth consumption path may be forced to choose activities with low variance in returns, possibly at the cost of lower mean. Over time, lower mean income leads to more poverty.

Credit and insurance markets are not the only area of the economy where limited information matters. Nor is lack of information the only constraint on economic interaction: The world we seek to explain is populated with economic actors who are boundedly rational, not rational. The fact that people are neither all-knowing nor have unlimited mental capability is important to us for several reasons.

One is that transactions become costly; and this problem is exacerbated as societies become larger and transactions more impersonal. Interaction with large societies requires more information about more people, which in turn requires

more calculation and processing (North 1993, 1995). Second, if we concede that agents are boundedly rational then we must distinguish between the objective world and each agent's subjective interpretation of the world. These interpretations are formed on the basis of individual and local experience, of individual inference and deduction, and of the intergenerational transmission of knowledge, values and customs. The product of these inputs is a mental model or belief system which drives, shapes and governs individual action (Simon 1986; North 1993).

These two implications of bounded rationality are important. The first (costly transactions) because when transactions are costly institutions matter. The second (local mental models and subjective beliefs) because these features of different countries and economies shape their institutions.

In this survey we emphasize two related aspects of institutions and their connection to poverty traps. The first is that institutions determine how well inefficiencies arising within the market are resolved. A typical example would be the efforts of economic and political institutions to solve coordination failure in a given activity resulting from some form of complementary externalities. The second is that institutions themselves can have inefficient equilibria. Moreover, institutions are *path dependent*. In the words of Paul A. David, they are the "carriers of history" (David 1994).

Why are institutions characterized by multiple equilibria and path dependence? Although human history often shows a pattern of negotiation towards efficient institutions which mitigate the cost of transactions and overcome market failure, it is also true that institutions are created and perpetuated by those with political power. As North (1993, p. 3) has emphasized, "institutions are not necessarily or even usually created to be socially efficient; rather they, or at least the formal rules, are created to serve the interests of those with the bargaining power to create new rules."

Moreover, the institutional framework is path dependent because those who currently hold power almost always have a stake in its perpetuation. Consider for example the current situation in Burundi, which has been mired in civil war since its first democratically elected president was assassinated in 1993. The economic consequences have not been efficient. Market-based economic activity has collapsed along with income. Life expectancy has fallen from 54 years in 1992 to 41 in 2000. Household final consumption expenditure is down 35% from 1980. Nevertheless, the military elite have much to gain from continuation of the war. The law of the gun benefits those with most guns. Curfew and

identity checks provide opportunities for extortion. Military leaders continue to subvert a peace process that would lead to reform of the army.

Path dependence is strengthened by positive feedback mechanisms which *reinforce* existing institutions. For example, the importance of strong property rights for growth has been extensively documented. Yet Acemoglu, Johnson and Robinson (this volume) document how in Europe during the Middle Ages monarchs consistently failed to ensure property rights for the general population. Instead they used arbitrary expropriation to increase their wealth and the wealth of their allies. Increased wealth closed the circle of causation by reinforcing their own power. Engerman and Sokoloff (2004) discuss how initial inequality in some of Europe's colonial possessions led to policies which hindered broad participation in market opportunities and strengthened the position of a small elite. Such policies tended to reinforce existing inequality (while acting as a break on economic growth).

Path dependence is also inherent in the way that informal norms form the foundations of community adherence to legal stipulations. While the legal framework can be changed almost instantaneously, social norms, conventions and other informal institutions are invariably persistent (otherwise they could hardly be conventions). Often legislation is just the first step a ruling body must take when seeking to alter the de facto rules of the game. [5]

Finally, bounded rationality can be a source of self-reinforcing inefficient outcomes independent of institutions. For example, even in an otherwise perfect market a lack of global knowledge can cause agents to choose an inefficient technology, which is then reinforced by herd effects. [6] When there are market frictions or nonconvexities such outcomes may be exacerbated. For example, if technology is nonconvex then initial poor choices by boundedly rational agents can be locked in (Arthur 1994).

In summary, the set of all self-reinforcing mechanisms which can potentially cause poverty is large. Even worse, the different mechanisms can interact, and reinforce one another. Increasing returns may cause investment complementarities and hence coordination failure, which is then perpetuated by pessimistic beliefs and conservative institutions. Rent-seeking and corruption may discour-

[5] For example, Transparency International's 2004 Global Corruption Report notes that in Zambia courts have been reluctant to hand down custodial sentences to those convicted of corruption, "principally because it was felt that white-collar criminals *did not deserve* to go to jail." (Emphasis added.)

[6] This example is due to Karla Hoff.

age investment in new technology, which lowers expected wages for skilled workers, decreasing education effort and hence the pool of skilled workers needed by firms investing in technology. The disaffected workers may turn to rent-seeking. Positive feedbacks reinforce other feedbacks. In these kinds of environments the relevance of the neoclassical benchmark seems tenuous at best.

Our survey of poverty traps proceeds as follows. Section 2 reviews key development facts. Section 3 considers several basic models associated with persistent poverty, and their implications for dynamics and the data. Section 4 looks at the empirics of poverty traps. Our survey of microfoundations is in Sections 5–8. Section 9 concludes.

There are already a number of surveys on poverty traps, including two by the first author (Azariadis 1996, 2004). The surveys by Hoff (2000) and Matsuyama (1995, 1997) are excellent, as is Easterly (2001). See in addition the edited volumes by Bowles, Durlauf and Hoff (2004) and Mookherjee and Ray (2001). Parente and Prescott (this volume) also focus on barriers to technology adoption as an explanation of cross-country variation in income levels. In their analysis institutions are treated as exogenous.

2 Development Facts

In Section 2.1 we briefly review key development facts, focusing on the vast and rising differences in per capita income across nations. Section 2.2 reminds the reader how these disparities came about by quickly surveying the economic history behind income divergence.

2.1 Poverty and riches

What does it mean to live on one or two dollars per day? Poverty translates into hunger, lack of shelter, illness without medical attention. Calorie intake in the poorest countries is far lower than in the rich. The malnourished are less productive and more susceptible to disease than those who are well fed. Infant mortality rates in the poorest countries are up to 40 or 50 times higher than the OECD average. Many of the common causes, such as pneumonia or dehydration from diarrhea, cost very little to treat.

The poor are more vulnerable to events they cannot control. They are less able to diversify their income sources. They are more likely to suffer from famine, violence and natural disasters. They have lower access to credit markets and insurance, with which to smooth out their consumption. Their children risk exploitation, and are less likely to become educated.

The plight of the poor is even more striking when compared to the remarkable wealth of the rich. Measured in 1996 US dollars and adjusted for purchasing power parity, average yearly income per capita in Luxembourg for 2000 was over $46,000.[7] In Tanzania, by contrast, average income for 2000 was about $500. In other words, people in Luxembourg are nearly 100 times richer on average than those living in the very poorest countries.[8] Luxembourg is rather exceptional in terms of per capita income, but even in the US average income is now about 70 times higher than it is in Tanzania.

How has the gap between the richest and the poorest evolved over time? The answer is simple: It has increased dramatically, even in the postwar era. In 1960, per capita income in Tanzania was $478. After rising somewhat during the 1960s and 1970s it collapsed again in the 1980s. By 2000 it was $457. Many other poor countries have had similar experiences, with income hovering around the $500–1,000 mark. Meanwhile, the rich countries continued exponential growth. Income in the US grew from $12,598 in 1960 (26 times that of Tanzania) to $33,523 in 2000 (73 times). Other rich industrialized countries had similar experiences. In Australia over the same period per capita GDP rose from $10,594 to $25,641. In France it rose from $7,998 to $22,253, and in Canada from $10,168 to $26,983.

Figure 1 shows how the rich have gotten richer relative to the poor. The left hand panel compares an average of real GDP per capita for the 5 richest countries in the Penn World Tables with an average of the same for the 5 poorest. The comparison is at each point in time from 1960 to the year 2000. The right panel does the same comparison with groups of 10 countries (10 richest vs 10 poorest) instead of 5. Both panels show that by these measures

[7] Unless otherwise stated, all income data in the remainder of this section is from the Penn World Tables Version 6.1 (Heston, Summers and Aten 2002). Units are PPP and terms of trade adjusted 1996 US dollars.

[8] Some countries record per capita income even lower than the figure given above for Tanzania. 1997 average income in Zaire is measured at $276. Sachs et al. (2004) use the World Bank's 2003 World Development Indicators to calculate a population-weighted average income for Sub-Saharan Africa at 267 PPP-adjusted US dollars, or 73 cents a day.

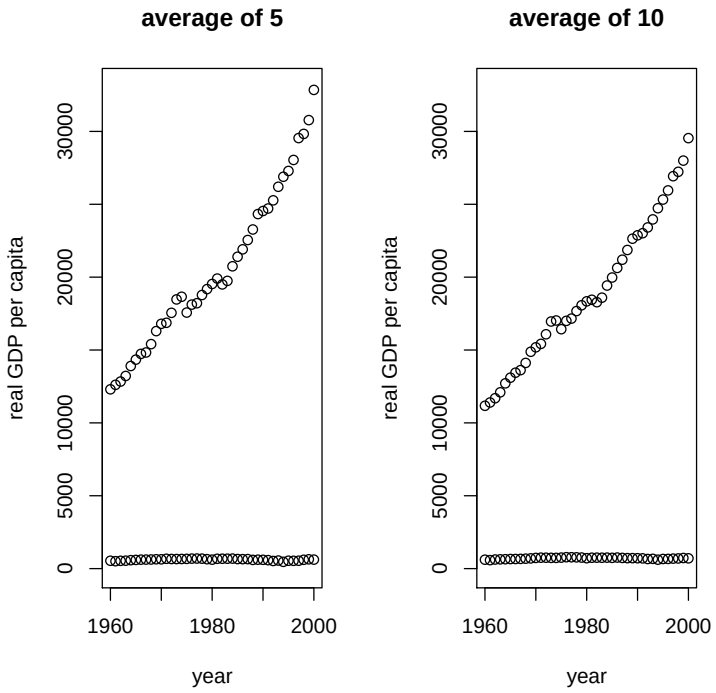


Fig. 1. The rich get richer

income disparity has widened dramatically in the postwar era, and the rate of divergence is, if anything, increasing. The vast and growing disparity in output per person shown in Figure 1 is what growth and development theorists are obliged to explain. [9]

2.2 A brief history of economic development

How did the massive disparities in income shown in Figure 1 arise? It is worth reviewing the broad history of economic development in order to remind ourselves of key facts. [10]

Although the beginnings of agriculture some ten thousand years ago marked the start of rapid human progress, for most of the subsequent millennia all but a tiny fraction of humanity was poor as we now define it, suffering regularly from hunger and highly vulnerable to adverse shocks. Early improvements

[9] Of course the figure says nothing about mobility. The poor this year could be the rich next year. See Section 4.1 for some discussion of mobility.

[10] The literature on origins of modern growth is too extensive to list here. See for example the monographs of Rostow (1975) and Mokyr (2002).

in economic welfare came with the rise of premodern city-states. Collective organization of irrigation, trade, communications and security proved more conducive to production than did autarky. Handicraft manufacture became more specialized over time, and agriculture more commercial. (Already the role of increasing returns and the importance of institutions are visible here.)

While such city-states and eventually large empires rose and fell over time, and the wealth of their citizens with them, until the last few hundred years no state successfully managed the transition to what we now call modern, self-sustaining growth. Increased wealth was followed by a rise in population. Malthusian pressure led to famine and disease.

The overriding reason for lack of sustained growth was that in the premodern world production technology improved only slowly. While the scientific achievements of the ancient Mediterranean civilizations and China were remarkable, in general there was little attempt to apply science to the economic problems of the peasants. Scientists and practical people had only limited interaction. Men and women of ability usual found that service to the state—or predation against other states—was more rewarding than entrepreneurship and invention.

Early signs of modern growth appeared in Western Europe around the middle of the last millennium. Science from the ancient world had been preserved, and now began to be extended. The revolutionary ideas of Copernicus led to intensive study of the natural world and its regularities. The printing press and movable type dramatically changed the way ideas were communicated. Innovations in navigation opened trade routes and new lands. Gunpowder and the cannon swept away local fiefdoms based on feudal castles.

These technological innovations led to changes in institutions. The weakening of local fiefdoms was followed in many countries by a consolidation of central authority, which increased the scale of markets and the scope for specialization.[11] Growing trade with the East and across the Atlantic produced a rich and powerful merchant class, who subsequently leveraged their political muscle to gain strengthened property and commercial rights.

Increases in market size, institutional reforms and progress in technology at

[11] For example, in 1664 Louis XIV of France drastically reduced local tolls and unified import customs. In 1707 England incorporated Scotland into its national market. Russia abolished internal duties in 1753, and the German states instituted similar reforms in 1808.

first lead to steady but unspectacular growth in incomes. In 1820 the richest countries in Europe had average per capita incomes of around $1,000 to $1,500—some two or three times that of the poorest countries today. However, in the early 19th Century the vast majority of people were still poor.

In this survey we compare productivity in the poor countries with the economic triumphs of the rich. Richness in our sense begins with the Industrial Revolution in Britain (although the rise in incomes was not immediate) and, subsequently, the rest of Western Europe. Industrialization—the systematic application of modern science to industrial technology and the rise of the factory system—led to productivity gains *entirely different in scale from those in the premodern world.*

In terms of proximate causes, the Industrial Revolution in Britain was driven by a remarkable revolution in science that occurred during the period from Copernicus through to Newton, and by what Mokyr (2002) has called the "Industrial Enlightenment," in which traditional artisanal practices were systematically surveyed, cataloged, analyzed and generalized by application of modern science. Critical to this process was the interactions of scientists with each other and with the inventors and practical men who sought to profit from innovation.

Science and invention led to breakthroughs in almost all areas of production; particularly transportation, communication and manufacturing. The structure of the British economy was massively transformed in a way that had never occurred before. Employment in agriculture fell from nearly 40% in 1820 to about 12% in 1913 (and to 2.2% in 1992). The stock of machinery, equipment and non-residential structures per worker increased by a factor of five between 1820 and 1890, and then doubled again by 1913. The literacy rate also climbed rapidly. Average years of education increased from 2 in 1820 to 4.4 in 1870 and 8.8 in 1913 (Maddison 1995).

As a result of these changes, per capita income in the UK jumped from about $1,700 in 1820 to $3,300 in 1870 and $5,000 in 1913. Other Western European countries followed suit. In the Netherlands, income per capita grew from $1,600 in 1820 to $4,000 in 1913, while for Germany the corresponding figures are $1,100 and $3,900. [12]

Looking forward from the start of the last century, it might have seemed likely that these riches would soon spread around the world. The innovations and

[12] The figures are from Maddison (1995). His units are 1990 international dollars.

inventions behind Britain's productivity miracle were to a large extent public knowledge. Clearly they were profitable. Adaptation to new environments is not costless, but nevertheless one suspects it was easy to feel that already the hard part had been done.

Such a forecast would have been far too optimistic. Relatively few countries besides Western Europe and its off-shoots have made the transition to modern growth. Much of the world remains mired in poverty. Among the worst performers are Sub-Saharan Africa and South Asia, which together account for some 70% of the 1.2 billion people living on less than $1 per day. But poverty rates are also high in East Asia, Latin America and the Carribean. Why is it that so many countries are still poorer than 19th Century Britain? Surely the different outcomes in Britain and a country such as Mali can—at least from a modeler's perspective—be Pareto ranked. What deviation from the neoclassical benchmark is it that causes technology growth in these countries to be retarded, and poverty to persist?

3 Models and Definitions

We begin our attempt to answer the question posed at the end of the last section with a review of the convex neoclassical growth model. It is appropriate to start with this model because it is the benchmark from which various deviations will be considered. Section 3.2 explains why the neoclassical model *cannot* explain the vast differences in income per capita between the rich and poor countries. Section 3.3 introduces the first of two "canonical" poverty trap models. These models allow us to address issues common to all such models, including dynamics and implications for the data. Section 3.4 introduces the second.

3.1 Neoclassical growth with diminishing returns

The convex neoclassical model (Solow 1956) begins with an aggregate production function of the form

$$Y_t = K_t^{\alpha}(A_t L_t)^{1-\alpha}\xi_{t+1}, \quad \alpha \in (0,1), \tag{1}$$

where Y is output of a single composite good, A is a productivity parameter, K is the aggregate stock of tangible and intangible capital, L is a measure of

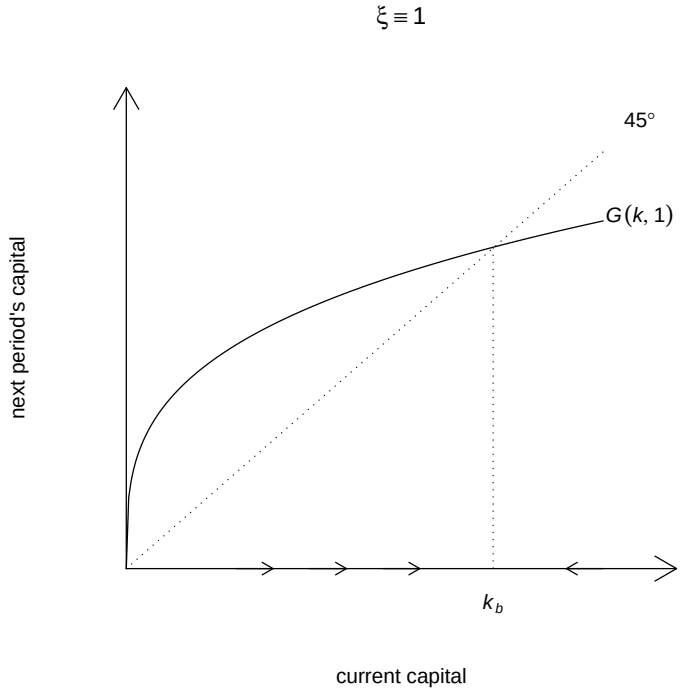


Fig. 2. Deterministic neoclassical dynamics

labor input, and ξ is a shock. In this formulation the sequence $(A_t)_{t\geq 0}$ captures the persistent component of productivity, and $(\xi_t)_{t\geq 0}$ is a serially uncorrelated innovation.

The production function on the right hand side of (1) represents maximum output for a given set of inputs. That output is maximal follows from competitive markets, profit seeking and free entry. (Implicit is the assumption of no significant indivisibilities or nonconvexities.) The Cobb-Douglas formulation is suggested by relative constancy of factor shares with respect to the level of worker output.

Savings of tangible and intangible capital from current output occurs at constant rate s; in which case K evolves according to the rule

$$K_{t+1} = sY_t + (1-\delta)K_t. \tag{2}$$

Here $\delta \in (0, 1]$ is a constant depreciation rate. The savings rate can be made endogenous by specifying intertemporal preferences. However the discussion in this section is purely qualitative; endogenizing savings changes little. [13]

[13] See, for example, Brock and Mirman (1972) or Nishimura and Stachurski (2004) for discussion of dynamics when savings is chosen optimally.

If, for example, labor L is undifferentiated and grows at exogenous rate n, and if productivity A is also exogenous and grows at rate γ, then the law of motion for capital per effective worker $k_t := K_t/(A_t L_t)$ is given by

$$k_{t+1} = \frac{s k_t^\alpha \xi_{t+1} + (1-\delta)k_t}{\theta} =: G(k_t, \xi_{t+1}), \tag{3}$$

where $\theta := 1 + n + \gamma$. The evolution of output per effective worker $Y_t/(A_t L_t)$ and output per capita Y_t/L_t are easily recovered from (1) and (3).

Because of diminishing returns, capital poor countries will extract greater marginal returns from each unit of capital stock invested than will countries with plenty of capital. The result is convergence to a long-run outcome which depends only on fundamental primitives (as opposed to beliefs, say, or historical conditions).

Figure 2 shows the usual deterministic global convergence result for this model when the shock ξ is suppressed. The steady state level of capital per effective worker is k_b. Figure 3 illustrates stochastic convergence with three simulated series from the law of motion (3), one with low initial income, one with medium initial income and one with high initial income. Part (a) of the figure gives the logarithm of output per effective worker, while (b) is the logarithm of output per worker. All three economies converge to the balanced growth path. [14]

Average convergence of the sample paths for $(k_t)_{t\geq 0}$ and income is mirrored by convergence in probabilistic laws. Consider for example the sequence of marginal distributions $(\psi_t)_{t\geq 0}$ corresponding to the sequence of random variables $(k_t)_{t\geq 0}$. Suppose for simplicity that the sequence of shocks is independent, identically distributed and lognormal; and that $k_0 > 0$. It can then be shown that (a) the distribution ψ_t is a density for all $t \geq 1$, and (b) the sequence $(\psi_t)_{t\geq 0}$ obeys the recursion

$$\psi_{t+1}(k') = \int_0^\infty \Gamma(k, k')\psi_t(k)dk, \quad \text{for all } t \geq 1, \tag{4}$$

where the *stochastic kernel* Γ in (4) has the interpretation that $\Gamma(k, \cdot)$ is the probability density for $k_{t+1} = G(k_t, \xi_{t+1})$ when k_t is taken as given and equal to k. [15] The interpretation of (4) is straightforward. It says (heuristically) that

[14] In the simulation the sequence of shocks $(\xi_t)_{t\geq 0}$ is lognormal, independent and identically distributed. The parameters are $\alpha = 0.3$, $A_0 = 100$, $\gamma = .025$, $n = 0$, $s = 0.2$, $\delta = 0.1$, and $\ln \xi \sim N(0, 0.1)$. Here and in all of what follows $X \sim N(\mu, \sigma)$ means that X is normally distributed with mean μ and standard deviation σ.

[15] See the technical appendix for details.

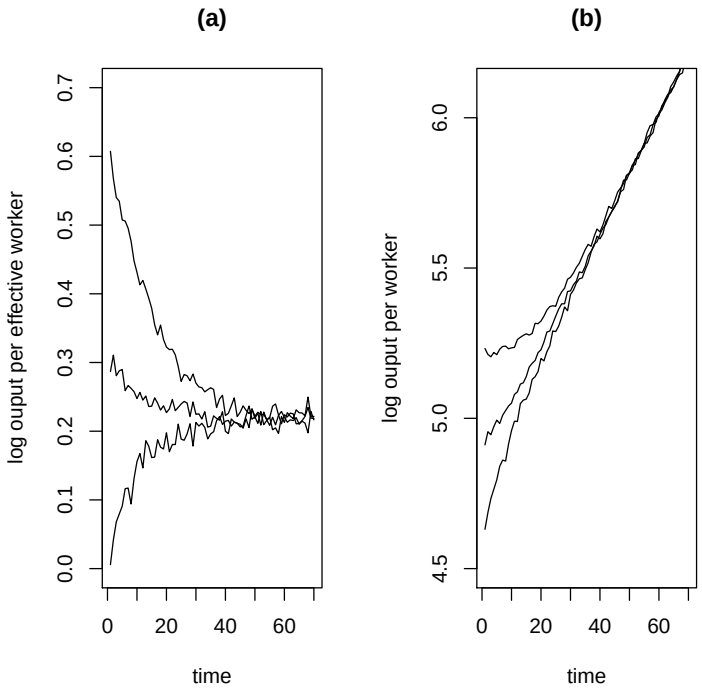


Fig. 3. Convergence to the balanced growth path

$\psi_{t+1}(k')$, the probability that k takes the value k' next period, is equal to the probability of taking value k' next period given that the current state is k, summed across all k, and weighted by $\psi_t(k)dk$, which is the probability that the current state actually takes the value k.

Here the conditional distribution $\Gamma(k, \cdot)$ of k_{t+1} given $k_t = k$ is easily calculated from (3) and the familiar change-of-variable rule that if ξ is a random variable with density φ and $Y = h(\xi)$, where h is smooth and strictly monotone, then Y has density $\varphi(h^{-1}(y)) \cdot [dh^{-1}(y)/dy]$. Applying this rule to (3) we get

$$\Gamma(k, k') := \varphi\left[\frac{\theta k' - (1-\delta)k}{sk^\alpha}\right]\frac{\theta}{sk^\alpha}, \tag{5}$$

where φ is the lognormal density of the productivity shock ξ. [16]

All Markov processes have the property that the sequences of marginal distributions they generate satisfies a recursion in the form of (4) for some stochastic kernel Γ. [17] Although the state variables usually do not themselves become

[16] Precisely, $z \mapsto \varphi(z)$ is this density when $z > 0$ and is equal to zero when $z \leq 0$.

[17] See the technical appendix for definitions. Note that we are working here with processes that generate sequences of *densities*. If the marginal distributions are not densities, and the conditional distribution contained in Γ is not a density, then the

stationary (due to the ongoing presence of noise), the sequence of probabilities $(\psi_t)_{t\geq 0}$ may. In particular, the following behavior is sometimes observed:

Definition 3.1 (Ergodicity) *Let a growth model be defined by some stochastic kernel* Γ*, and let* $(\psi_t)_{t\geq 0}$ *be the corresponding sequence of marginal distributions generated by* (4)*. The model is called* ergodic *if there is a unique probability distribution* ψ^* *supported on* $(0, \infty)$ *with the property that* (i)

$$\psi^*(k') = \int_0^\infty \Gamma(k, k')\psi^*(k)dk \quad \textit{for all } k';$$

and (ii) *the sequence* $(\psi_t)_{t\geq 0}$ *of marginal distributions for the state variable satisfies* $\psi_t \to \psi^*$ *as* $t \to \infty$ *for all non-zero initial states.* [18]

It is easy to see that (i) and (4) together imply that if $\psi_t = \psi^*$ (that is, $k_t \sim \psi^*$), then $\psi_{t+1} = \psi^*$ (that is, $k_{t+1} \sim \psi^*$) also holds (and if this is the case then $k_{t+2} \sim \psi^*$ follows, and so on). A distribution with this property is called a stationary distribution, or ergodic distribution, for the Markov chain. Property (ii) says that, conditional on a strictly positive initial stock of capital, the marginal distribution of the stock converges in the long run to the ergodic distribution.

Under the current assumptions it is relatively straightforward to prove that the Solow process (3) is ergodic. (See the technical appendix for more details.) Figures 4 and 5 show convergence in the neoclassical model (3) to the ergodic distribution ψ^*. In each of the two figures an initial distribution ψ_0 has been chosen arbitrarily. Since the process is ergodic, in both figures the sequence of marginal distributions $(\psi_t)_{t\geq 0}$ converges to the same ergodic distribution ψ^*. This distribution ψ^* is determined purely by fundamentals, such as the propensity to save, the rate of capital depreciation and fertility. [19]

---

formula (4) needs to be modified accordingly. See the technical appendix. Other references include Stokey, Lucas and Prescott (1989), Futia (1982) and Stachurski (2004).

[18] Convergence refers here to that of measures in the total variation norm, which in this case is just the L_1 norm. Convergence in the norm topology implies convergence in distribution in the usual sense.

[19] The algorithms and code for computing marginal and ergodic distributions are available from the authors. All ergodic distributions are calculated using Glynn and Henderson's (2001) look-ahead estimator. Marginals are calculated using a variation of this estimator constructed by the authors. The parameters in (3) are chosen—rather arbitrarily—as $\alpha = 0.3$, $\gamma = .02$, $n = 0$, $s = 0.2$, $\delta = 1$, and $\ln \xi \sim N(3.6, 0.11)$.

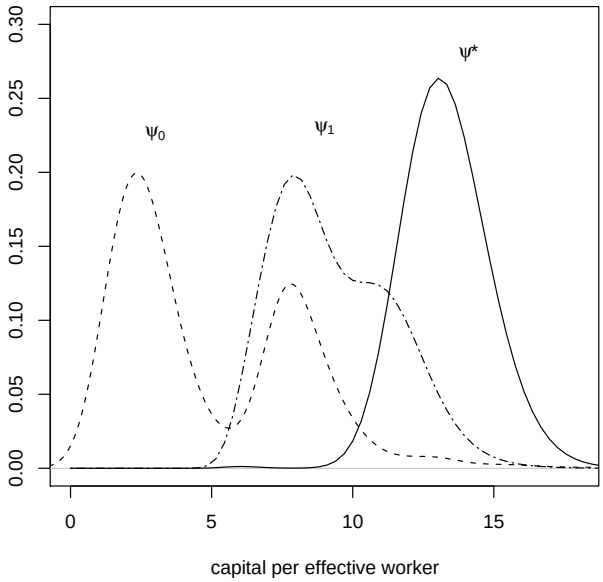


Fig. 4. Convergence to the ergodic distribution

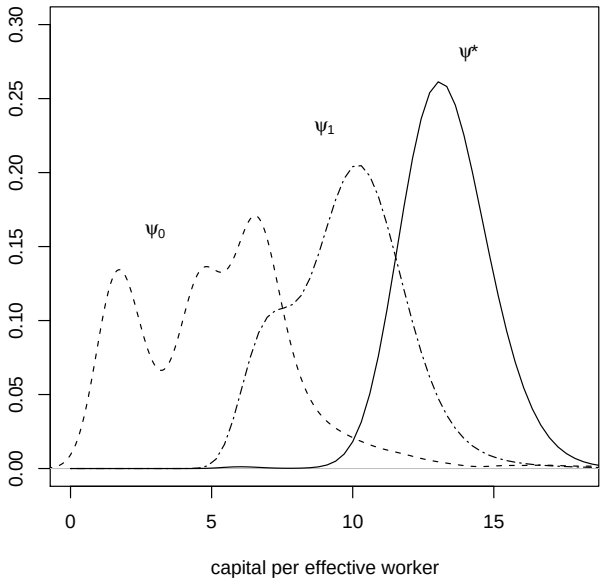


Fig. 5. Convergence to the ergodic distribution

Notice in Figures 4 and 5 how initial differences are moderated under the convex neoclassical transition rule. We will see that, without convexity, initial differences often persist, and may well be amplified as the system evolves through time.

The convex neoclassical growth model described in the previous section predicts that per capita incomes will differ across countries with different rates of physical and human capital formation or fertility. Can the model provide a reasonable explanation then for the fact that per capita income in the US is more than 70 times that in Tanzania or Malawi?

The short answer to this question is no. First, rates at which people accumulate reproducible factors of production or have children (fertility rates) are endogenous—in fact they are choice variables. To the extent that factor accumulation and fertility are important, we need to know *why* some individuals and societies make choices that lead them into poverty. For poverty is suffering, and, all things being equal, few people will choose it.

This same observation leads us to suspect that the choices facing individuals in rich countries and those facing individuals in poor countries are very different. In poor countries, the choices that collectively would drive modern growth—innovation, investment in human and physical capital, etc.—must be perceived by individuals as *worse* than those which collectively lead to the status quo. [20]

A second problem for the convex neoclassical growth model as an explanation of level differences is that even when we regard accumulation and fertility rates as exogenous, they must still account for all variation in income per capita across countries. However, as many economists have pointed out, the differences in savings and fertility rates are not large enough to explain real income per capita ratios in the neighborhood of 70 or 100. A model ascribing output variation to these few attributes alone is insufficient. A cotton farmer in the US does not produce more cotton than a cotton farmer in Mali simply because he has saved more cotton seed. The production techniques used in these two countries are utterly different, from land clearing to furrowing to planting to irrigation and to harvest. A model which does not address the vast differences in production technology across countries cannot explain the observed differences in output.

Let us very briefly review the quantitative version of this argument. [21] To

[20] For this reason, endogenizing savings by specifying preferences is not very helpful, because to get poverty in optimal growth models we must assume that the poor are poor because they prefer poverty.

[21] The review is brief because there are many good sources. See, for example, Lucas (1990), King and Rebelo (1993), Prescott (1998), Hall and Jones (1999) or Easterly

begin, recall the aggregate production function (1), which is repeated here for convenience:

$$Y_t = K_t^{\alpha}(A_t L_t)^{1-\alpha}\xi_{t+1}. \tag{6}$$

All of the components are more or less observable besides A_t and the shock.[22] Hall and Jones (1999) conducted a simple growth accounting study by collecting data on the observable components for the year 1988. They calculate that the geometric average of output per worker for the 5 richest countries in their sample was 31.7 times that of the 5 poorest countries. Taking L to be a measure of human capital, variation in the two inputs L and K contributed only factors of 2.2 and 1.8 respectively. This leaves all the remaining variation in the productivity term A.[23]

This is not a promising start for the neoclassical model as a theory of level differences. Essentially, it says that there is no single map from total inputs to aggregate output that holds for every country. Why might this be the case? We know that the aggregate production function is based on a great deal of theory. Output is maximal for a given set of inputs because of perfect competition among firms. Free entry, convex technology relative to market size, price taking and profit maximization mean that the best technologies are used—and used efficiently. Clearly some aspect of this theory must deviate significantly from reality.

Now consider how this translates into predictions about level differences in income per capita. When the shock is suppressed ($\xi_t = 1$ for all t), output per capita converges to the balanced path

$$y_t := \frac{Y_t}{L_t} = A_t(s/\kappa)^{\alpha/(1-\alpha)}, \tag{7}$$

where $\kappa := n + \gamma + \delta$.[24] Suppose at first that the path for the productivity residual is the same in all countries. That is, $A_t^i = A_t^j$ for all i, j and t. In this

and Levine (2000).

[22] The parameter α is the share of capital in the national accounts. Human capital can be estimated by collecting data on total labor input, schooling, and returns to each year of schooling as a measure of its productivity.

[23] The domestic production shocks $(\xi_t)_{t\geq 0}$ are not the source of the variation. This is because they are very small relative to the differences in incomes across countries, and, by definition, not persistent. (Recall that in our model they are innovations to the permanent component $(A_t)_{t\geq 0}$.)

[24] When considering income *levels* it is necessary to assume that countries are in the neighborhood of the balanced path, for this is where the model predicts they will be. Permitting them to be "somewhere else" is not a theory of variations in income levels.

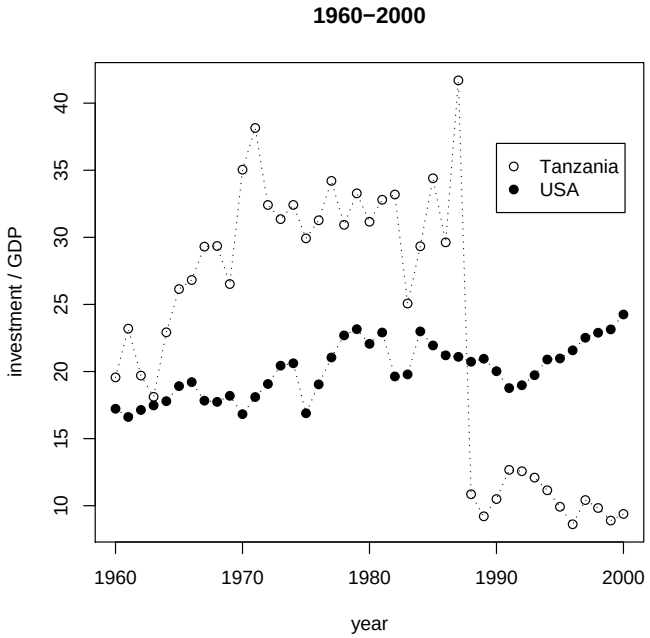


Fig. 6. Investment rates in Tanzania and the US

case, the ratio of output per capita in country i relative to that in country j is constant and equal to

$$\frac{y^i}{y^j} = \left(\frac{s^i \kappa^j}{s^j \kappa^i}\right)^{\alpha/(1-\alpha)}. \tag{8}$$

The problem for the neoclassical model is that the term inside the brackets is usually not very large. For example, if we compare the US and Tanzania, say, and if we identify capital with physical capital, then average investment as a fraction of GDP between 1960 and 2000 was about 0.2 in the US and 0.24 in Tanzania. (Although the rate in Tanzania varied a great deal around this average. See Figure 6.) The average population growth rates over this period were about 0.01 and 0.03 respectively. Since $A_t^i = A_t^j$ for all t we have $\gamma^i = \gamma^j$. Suppose that this rate is 0.02, say, and that $\delta^i = \delta^j = 0.05$. This gives $s^i \kappa^j / (s^j \kappa^i) \approx 1$. Since payments to factors of production suggest that $\alpha/(1-\alpha)$ is neither very large nor very small, output per worker in the two countries is predicted to be roughly equal.

This is only an elementary calculation. The computation of investment rates in Tanzania is not very reliable. There are issues in terms of the relative ratios of consumption and investment good prices in the two countries which may distort the data. Further, we have not included intangible capital—most notably human capital. The rate of investment in human capital and training

in the US is larger than it is in Tanzania. Nevertheless, it is difficult to get the term in (8) to contribute a factor of much more than 4 or 5—certainly not 70.[25]

However the calculations are performed, it turns out that to explain the ratio of incomes in countries such as Tanzania and the US, productivity residuals must absorb most of the variation. In other words, the convex neoclassical growth model cannot be reconciled with the cross-country income data *unless we leave most of the variation in income to an unexplained residual term about which we have no quantitative theory.* And surely *any* scientific theory can explain *any* given phenomenon by adopting such a strategy.

Different authors have made this same point in different ways. Lucas (1990) points out that if factor input differences *are* large enough to explain cross-country variations in income, the returns to investment in physical and human capital in poor countries implied by the model will be huge compared to those found in the rich. In fact they are not. Also, productivity residuals are growing quickly in countries like the US.[26] On the other hand, in countries like Tanzania, growth in the productivity residual has been very small.[27] Yet the convex neoclassical model provides no theory on why these different rates of growth in productivity should hold.

On balance, the importance of productivity residuals suggests that the poor countries are not rich because for one reason or another they have failed or not been able to adopt modern techniques of production. In fact production technology in the poorest countries is barely changing. In West Africa, for example, almost 100% of the increase in per capita food output since 1960 has come from expansion of harvest area (Baker 2004). On the other hand, the rich countries are becoming ever richer because of continued innovation.

[25] See in particular Prescott (1998) for detailed calculations. He concludes that convex neoclassical growth theory "fails as a theory of international income differences, even after the concept of capital is broadened to include human and other forms of intangible capital. It fails because differences in savings rates cannot account for the great disparity in per capita incomes unless investment in intangible capital is implausibly large."

[26] One can compute this directly, or infer it from the fact that interest rates in the US have shown no secular trend over the last century, in which case transitional dynamics can explain little, and therefore growth in output per worker and growth in the residual can be closely identified (King and Rebelo 1993).

[27] Again, this can be computed directly, or inferred from the fact that if it had been growing at a rate similar to the US, then income in Tanzania would have been at impossibly low levels in the recent past (Pritchett 1997).

Of course this only pushes the question one step back. Technological change is only a proximate cause of diverging incomes. What economists need to explain is why production technology has improved so quickly in the US or Japan, say, and comparatively little in countries such as Tanzania, Mali and Senegal.

We end this section with some caveats. First, the failure of the simple convex neoclassical model does *not* imply the existence of poverty traps. For example, we may discover successful theories that predict very low levels of the residual based on exogenous features which tend to characterize poor countries. (Although it may turn out that, depending on what one is prepared to call exogenous, the map from fundamentals to outcomes is not uniquely defined. In other words, there are multiple equilibria. In Section 4.2 some evidence is presented on this point.)

Further, none of the discussion in this section seeks to deny that factor accumulation matters. Low rates of factor accumulation are certainly correlated with poor performance, and we do not wish to enter the "factor accumulation versus technology" debate—partly because this is viewed as a contest between neoclassical and "endogenous" growth models, which is tangential to our interests, and partly because technology and factor accumulation are clearly interrelated: technology drives capital formation and investment boosts productivity. [28]

Finally, it should be emphasized that our ability to reject the elementary convex neoclassical growth model as a theory of level differences between rich and poor countries is precisely because of its firm foundations in theory and excellent quantitative properties. All of the poverty trap models we present in this survey provide far less in terms of quantitative, testable restrictions that can be confronted with the data. The power of a model depends on its falsifiability, not its potential to account for every data set.

3.3 Poverty traps: historical self-reinforcement

How then are we to explain the great variation in cross-country incomes such as shown in Figure 1? In the introduction we discussed some deviations from the neoclassical benchmark which can potentially account for this variation by endogenously reinforcing small initial differences. Before going into the

[28] However, as we stressed at the beginning of this section, to the extent that factor accumulation is important it may in fact turn out that low accumulation rates are mere symptoms of poverty, not causes.

specifics of different feedback mechanisms, this section formulates the first of two abstract poverty trap models. For both models a detailed investigation of microfoundations is omitted. Instead, our purpose is to establish a framework for the questions poverty traps raise about dynamics, and for their observable implications in terms of the cross-country income data.

The first model—a variation on the convex neoclassical growth model discussed in Section 3.1—is loosely based on Romer (1986) and Azariadis and Drazen (1990). It exemplifies what Mookherjee and Ray (2001) have called *historical* self-reinforcement, a process whereby initial conditions of the endogenous variables can shape long run outcomes. Leaving aside all serious complications for the moment, let us fix at $s > 0$ the savings rate, and at zero the rates of exogenous technological progress γ and population growth n. Let all labor be undifferentiated and normalize its total mass to 1, so that k represents both aggregate capital and capital per worker. Suppose that the productivity parameter A can vary with the stock of capital. In other words, A is a function of k, and aggregate returns $k_t \mapsto A(k_t)k_t^\alpha$ are potentially increasing. [29]

The law of motion for the economy is then

$$k_{t+1} = sA(k_t)k_t^\alpha \xi_{t+1} + (1-\delta)k_t. \tag{9}$$

Depending on the specification of the relationship between k and productivity, many dynamic paths are possible. Some of them will lead to poverty traps. Figure 7 gives examples of potential dynamic structures. For now the shock ξ is suppressed. The x-axis is current capital k_t and the y-axis is k_{t+1}. In each case the plotted curve is just the right hand side of (9), all with different maps $k \mapsto A(k)$.

In part (a) of the figure the main feature is non-ergodic dynamics: long run outcomes depend on the initial condition. Specifically, there are two local attractors, the basins of attraction for which are delineated by the unstable fixed point k_b. Part (b) is also non-ergodic. It shows the same low level attractor, but now no high level attractor exists. Beginning at a state above k_b leads to unbounded growth. In part (c) the low level attractor is at zero.

[29] In Romer (1986), for example, private investment generates new knowledge, some of which enters the public domain and can be used by other firms. In Azariadis and Drazen (1990) there are spillovers from human capital formation. See also Durlauf (1993) and Zilibotti (1995). See Matsuyama (1997) and references for discussion of how investment may feed back via pecuniary externalities into specialization and hence productivity. Our discussion of microfoundations begins in Section 5.

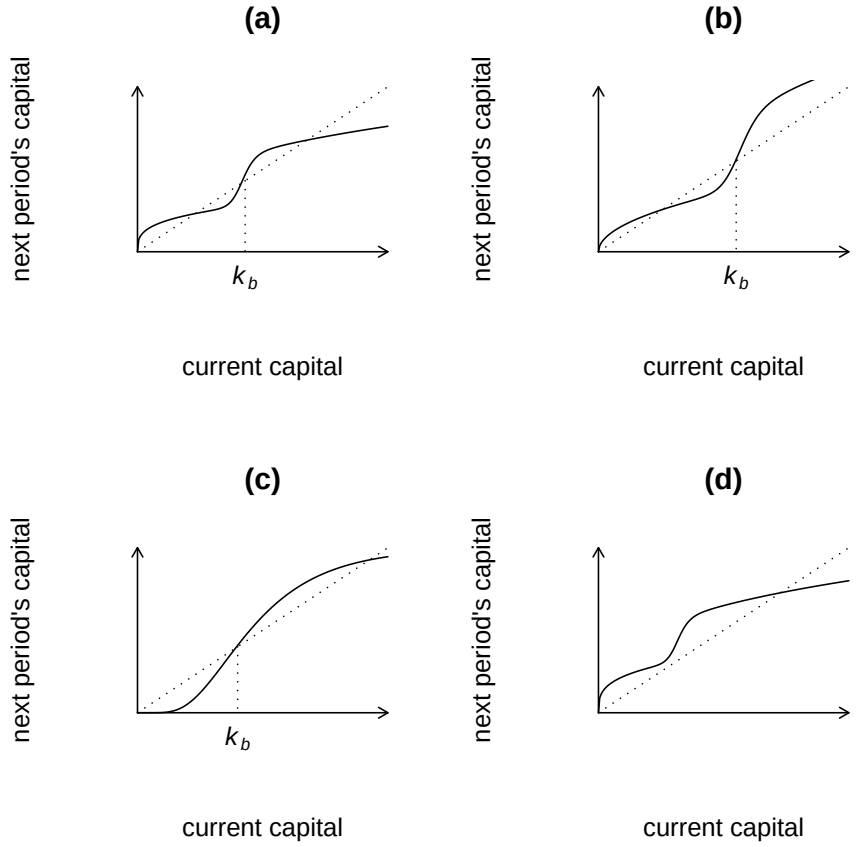


Fig. 7. Models with poverty traps

The figure in part (d) looks like an anomaly. Since the dynamics are formally ergodic, many researchers will not view this structure as a "poverty trap" model. Below we argue that this reading is too hasty: the model in (d) can certainly generate the kind of persistent-poverty aggregate income data we are hoping to explain.

In order to gain a more sophisticated understanding, let us now look at the stochastic dynamics of the capital stock. Deterministic dynamics are of course a special case of stochastic dynamics (with zero-variance shocks) but as in the case of the neoclassical model above, let us suppose that $(\xi_t)_{t\geq 0}$ is independently and identically lognormally distributed, with $\ln \xi \sim N(\mu, \sigma)$ and $\sigma > 0$. It then follows that the sequence of marginal distributions $(\psi_t)_{t\geq 0}$ for the capital stock sequence $(k_t)_{t\geq 0}$ again obeys the recursion (4) where the stochastic kernel Γ is now

$$\Gamma(k, k') := \varphi\left[\frac{k' - (1-\delta)k}{sA(k)k^\alpha}\right] \frac{1}{sA(k)k^\alpha}, \tag{10}$$

with φ the lognormal density on $(0, \infty)$ and zero elsewhere. All of the intuition for the recursion (4) and the construction of the stochastic kernel (10) is exactly the same as the neoclassical case.

How do the marginal distributions of the nonconvex growth model evolve? The following result gives the answer for most cases we are interested in.

Proposition 3.1 *Let $(\xi_t)_{t\geq 0}$ be an independent sequence with $\ln \xi_t \sim N(\mu, \sigma)$ for all t. If the function $k \mapsto A(k)$ satisfies the regularity condition*

$$0 < \inf_k A(k) \leq \sup_k A(k) < \infty,$$

then the stochastic nonconvex growth model defined by (9) *is ergodic.* [30]

Ergodicity here refers to Definition 3.1 on page 18, which, incidentally, is the standard definition used in growth theory and macroeconomics (see, for example, Brock and Mirman 1972; or Stokey, Lucas and Prescott 1989). In other words, there is a unique ergodic distribution ψ^*, and the sequence of marginal distributions $(\psi_t)_{t\geq 0}$ converges to ψ^* asymptotically, independent of the initial condition (assuming of course that $k_0 > 0$). A proof of this result is given in the technical appendix.

So why has a non-ergodic model become ergodic with the introduction of noise? The intuition is completely straightforward: Under our assumption of unbounded shocks there is always the potential—however small—to escape any basin of attraction. So in the long run initial conditions do not matter. (What *does* matter is how long this long run is, a point we will return to below.)

Figure 8 gives the ergodic distributions corresponding to two poverty trap models. [31] Both have the same structural dynamics as the model in part (a) of Figure 7. The left hand panels show this structure with the shock suppressed. The right hand panels show corresponding ergodic distributions under the independent lognormal shock process. Both ergodic distributions are bimodal, with modes concentrated around the deterministic local attractors.

Comparing the two left hand panels, notice that although qualitatively similar,

[30] In fact we require also that $k \mapsto A(k)$ is a Borel measurable function. But this condition is very weak indeed. For example, $k \mapsto A(k)$ need be neither monotone nor continuous.

[31] Regarding numerical computation see the discussion for the neoclassical case above.

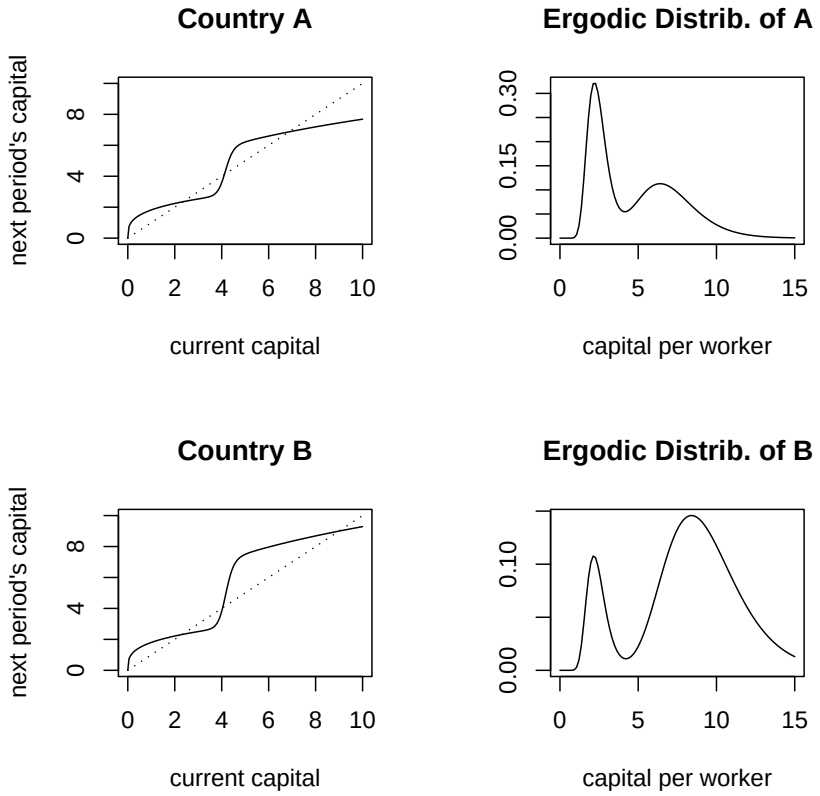


Fig. 8. Ergodic distributions under increasing returns

the laws of motion for Country A and Country B have different degrees of increasing returns. For Country B, the jump occurring around $k = 4$ is larger. As a result, the state is less likely to return to the neighborhood of the lower attractor once it makes the transition out of the poverty trap. Therefore the mode of the ergodic distribution corresponding to the higher attractor is large relative to that of Country A. Economies driven by law of motion B spend more time being rich.

Convergence to the ergodic distribution in a nonconvex growth model is illustrated in Figure 9. The underlying model is (a) of Figure 7.[32] As before, the ergodic distribution is bimodal. In this simulation, the initial distribution was chosen arbitrarily. Note how initial differences tend to be magnified over the medium term despite ergodicity. The initially rich diverge to the higher

[32] The specification of $A(k)$ used in the simulation is $A(k) = a \exp(h\Psi(k))$, where $a = 15$, $h = 0.52$ and the transition function Ψ is given by $\Psi(k) := (1 + \exp(-\ln(k/k_T)/\theta))^{-1}$. The parameter k_T is a "threshold" value of k, and is set at 6.9. The parameter θ is the smoothness of the transition, and is set at 0.09. The other parameters are $\alpha = 0.3$, $s = 0.2$, $\delta = 1$, and $\ln \xi \sim N(0, 0.1)$.

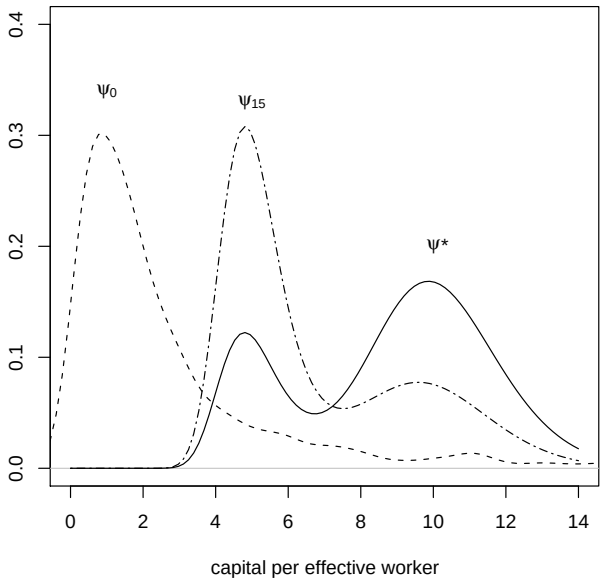


Fig. 9. Convergence under increasing returns

mode, creating the kind of "convergence club" effect seen in ψ_{15}, the period 15 marginal distribution. [33]

It is clear, therefore, that ergodicity is not the whole story. If the support of the shock ξ is bounded then ergodicity may not hold. Moreover, even with ergodicity, historical conditions may be arbitrarily persistent. Just how long they persist depends mainly on (i) the size of the basins of attraction and (ii) the statistical properties of the shock. On the other hand, the non-zero degree of mixing across the state space that drives ergodicity is usually more realistic than deterministic models where poverty traps are absolute and can never be overcome. Indeed, we will see that ergodicity is very useful for framing empirical questions in Section 4.2.

Figures 10 and 11 illustrate how historical conditions persist for individual time series generated by a model in the form of (a) of Figure 7, regardless of ergodicity. In both figures, the x-axis is time and the y-axis is (the log of) capital stock per worker. The dashed line through the middle of the figure corresponds to (the log of) k_b, the point dividing the two basins of attraction in (a) of Figure 7. Both figures show the simulated time series of four

[33] Incidentally, the change in the distributions from ψ_0 to ψ_{15} is qualitatively quite similar to the change in the cross-country income distribution that has been observed in the post war period.

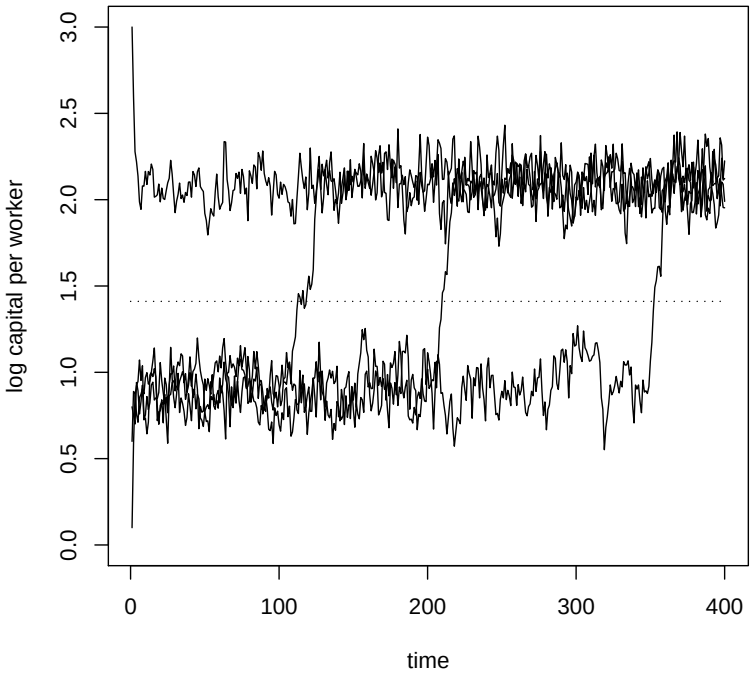


Fig. 10. Time series of 4 countries, high variance

economies. In each figure, all four economies are identical, apart from their initial conditions. One economy is started in the basin of attraction for the higher attractor, and three are started in that of the lower attractor. [34]

In the figures, the economies spend most of the time clustered in the neighborhoods of the two deterministic attractors. Economies starting in the portion of the state space (the y-axis) above the threshold are attracted on average to the high level attractor, while those starting below are attracted on average to the low level attractor. For these parameters, historical conditions are important in determining outcomes over the kinds of time scales economists are interested in, even though there are no multiple equilibria, and in the limit outcomes depend only on fundamentals.

In Figure 10, all three initially poor economies eventually make the transition out of the poverty trap, and converge to the neighborhood of the high attractor. Such transitions might be referred to as "growth miracles." In these

[34] The specification of $A(k)$ is as in Figure 9, where now $k_T = 4.1$, $\theta = 0.2$, $h = 0.95$, $\alpha = 0.3$, $s = 0.2$, $\delta = 1$. For Figure 10 we used $\ln \xi \sim N(0, 0.1)$, while for Figure 11 we used $\ln \xi \sim N(0, 0.05)$.

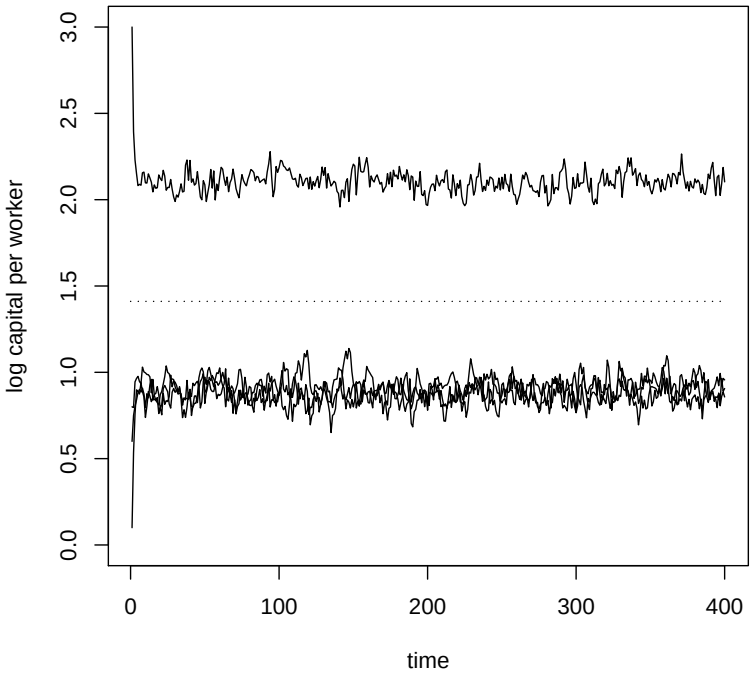


Fig. 11. Time series of 4 countries, low variance

series there are no "growth disasters" (transitions from high to low). The relative likelihood of growth miracles and growth disasters obviously depends on the structure of the model—in particular, on the relative size of the basins of attraction.

In Figure 10 the shock is distributed according to $\ln \xi \sim N(0, 0.1)$, while in Figure 11 the variance is smaller: $\ln \xi \sim N(0, 0.05)$. Notice that in Figure 11 no growth miracles occur over this time period. The intuition is clear: With less noise, the probability of a large positive shock—large enough to move into the basin of attraction for the high attractor—is reduced, and with it the probability of escaping from the poverty trap.

We now return to the model in part (d) of Figure 7, which is nonconvex, but at the same time is ergodic even in the deterministic case. This kind of structure is usually *not* regarded as a poverty trap model. In fact, since (d) is just a small perturbation of model (a), the existence of poverty traps is often thought to be very sensitive to parameters—a small change can cause a bifurcation of the dynamics whereby the poverty trap disappears. But, in fact, the phenomenon of persistence is more subtle. In terms of their medium run implications for cross-country income patterns, the two models (a) and

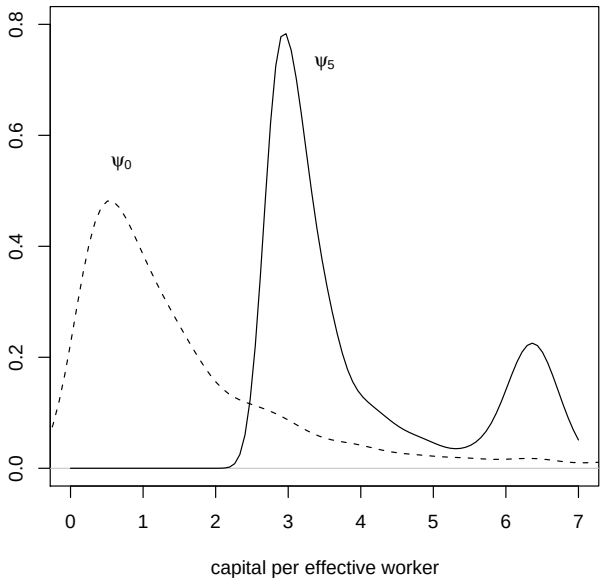


Fig. 12. Convergence under increasing returns

(d) are very similar.

To illustrate this, Figure 12 shows an arbitrary initial distribution and the resulting time 5 distribution for k under the law of motion given in (d) of Figure 7.[35] As in all cases we have considered, the stochastic model is ergodic. Now the ergodic distribution (not shown) is unimodal, clustered around the single high level attractor of the deterministic model. Thus the long run dynamics are different to those in Figure 9. *However,* during the transition, statistical behavior is qualitatively the same as that for models that *do* have low level attractors (such as (a) of Figure 7). In ψ_5 we observe amplification of initial differences, and the formation of a bimodal distribution with two "convergence clubs."

How long is the medium run, when the transition is in progress and the distribution is bimodal? In fact one can make this transition arbitrarily long without changing the basic qualitative features of (d), such as the non-existence of a low level attractor. Its length depends on the degree of nonconvexity and the variance of the productivity shocks $(\xi_t)_{t\geq 0}$. Higher variance in the shocks will tend to speed up the transition.

[35] The specification of $A(k)$ is as before, where now $k_T = 3.1$, $\theta = 0.15$, $h = 0.7$, $\alpha = 0.3$, $s = 0.2$, $\delta = 1$, and $\ln \xi \sim N(0, 0.2)$.

The last two examples have illustrated an important general principle: In economies with nonconvexities, the dynamics of key variables such as income can be highly sensitive to the statistical properties of the exogenous shocks which perturb activity in each period. [36] This phenomenon is consistent with the cross-country income panel. Indeed, several studies have emphasized the major role that shocks play in determining the time path of economic development (c.f., e.g., Easterly, Kremer, Pritchett and Summers 1993; den Haan 1995; Acemoglu and Zilibotti 1997; Easterly and Levine 2000). [37]

At the risk of some redundancy, let us end our discussion of the increasing returns model (9) by reiterating that persistence of historical conditions and formal ergodicity may easily coincide. (Recall that the time series in Figure 11 are generated by an ergodic model, and that (d) of Figure 7 is ergodic even in the deterministic case.) As a result, identifying history dependence with a lack of ergodicity can be problematic. In this survey we use a more general definition:

Definition 3.2 (Poverty trap) *A poverty trap is any self-reinforcing mechanism which causes poverty to persist.*

When considering a given quantitative model and its dynamic implications, the important question to address is, how persistent are the self-reinforcing mechanisms which serve to lock in poverty over the time scales that matter when welfare is computed? [38]

A final point regarding this definition is that the mechanisms which reinforce poverty may occur at any scale of social and spatial aggregation, from individuals to families, communities, regions, and countries. Traps can arise not just across geographical location such as national boundaries, but also within dispersed collections of individuals affiliated by ethnicity, religious beliefs or clan. Group outcomes are then summed up progressively from the level of the individual. [39]

[36] Such sensitivity is common to all dynamic systems where feedbacks can be positive. The classic example is evolutionary selection.

[37] This point also illustrates a problem with standard empirical growth studies. In general no information on the shock distribution is incorporated into calculation of dynamics.

[38] Mookherjee and Ray (2001) have emphasized the same point. See their discussion of "self-reinforcement as slow convergence."

[39] This point has been emphasized by Barrett and Swallow (2003) in their discussion of "fractal" poverty traps.

Next we turn to our second "canonical" poverty trap model, which again is presented in a very simplistic form. (For microfoundations see Sections 5–8.) The model is static rather than dynamic, and exhibits what Mookherjee and Ray (2001) have described as *inertial* self-reinforcement. [40] Multiple equilibria exist, and selection of a particular equilibrium can be determined purely by beliefs or subjective expectations.

In the economy a unit mass of agents choose to work either in a traditional, rural sector or a modern sector. Labor is the only input to production, and each agent supplies one unit in every period. All markets are competitive. In the traditional sector returns to scale are constant, and output per worker is normalized to zero. The modern sector, however, is knowledge-intensive, and aggregate output exhibits increasing returns due perhaps to spillovers from agglomeration, or from matching and network effects.

Let the fraction of agents working in the modern sector be denoted by α. The map $\alpha \mapsto f(\alpha)$ gives output per worker in the modern sector as a function of the fraction employed there. Payoffs are just wages, which equal output per worker (marginal product). Agents maximize individual payoffs taking the share α as exogenously given.

We are particularly interested in the case of strategic complementarities. Here, entry into the modern sector exhibits complementarities if the payoff to entering the modern sector increases with the number of other agents already there; in other words, if f is increasing. We assume that $f' > 0$, and also that returns in the modern sector dominate those in the traditional sector only when the number of agents in the modern sector rises above some threshold. That is, $f(0) < 0 < f(1)$. This situation is shown in Figure 13. At the point α_b returns in the two sectors are equal.

Equilibrium distributions of agents are values of α such that $f(\alpha) = 0$, as well as "all workers are in the traditional sector," or "all workers are in the modern sector" (ignoring adjustments on null sets). The last two of these are clearly Pareto-ranked: The equilibrium $\alpha = 0$ has the interpretation of a poverty trap.

Immediately the following objection arises. Although the lower equilibrium is to be called a poverty trap, is there really a self-reinforcing mechanism here

[40] By "static" we mean that there are no explicitly specified interactions between separate periods.

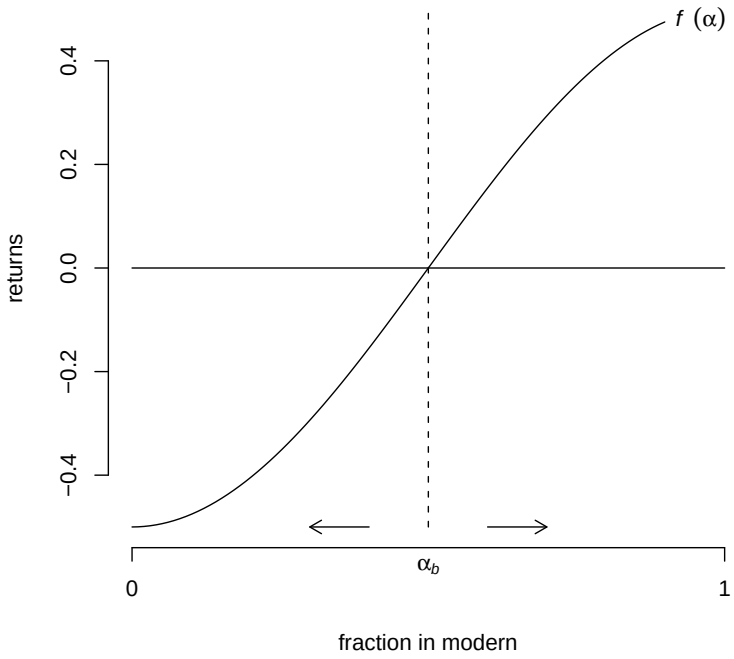


Fig. 13. Returns in traditional and modern sectors

which causes poverty to persist? After all, it seems that as soon as agents coordinate on the good equilibrium "poverty" will disappear. And there are plenty of occasions where societies acting collectively have put in place the institutions and preconditions for successful coordination when it is profitable to do so.

Although the last statement is true, it seems that history still has a role to play in equilibrium selection. This argument has been discussed at some length in the literature, usually beginning with myopic Marshallian dynamics, under which factors of production move over time towards activities where returns are higher. In the case of our model, these dynamics are given by the arrows in Figure 13. If $(\alpha_0)_{t\geq 0}$ is the sequence of modern sector shares, and if initially $\alpha_0 < \alpha_b$, then $\alpha_t \to 0$. Conversely, if $\alpha_0 > \alpha_b$, then $\alpha_t \to 1$.

But, as many authors have noted, this analysis only pushes the question one step back. Why should the sectoral shares only evolve slowly? And if they can adjust instantaneously, then why should they depend on the initial condition at all? What are the sources of inertia here that prevent agents from immediately

coordinating on good equilibria? [41]

Adsera and Ray (1997) have proposed one answer. Historical conditions may be decisive if—as seems quite plausible—spillovers in the modern sector arise only with a lag. A simplified version of the argument is as follows. Suppose that the private return to working in the modern sector is r_t, where now $r_0 = f(\alpha_0)$ and r_t takes the lagged value $f(\alpha_{t-1})$ when $t \geq 1$. Suppose also that at the end of each period agents can move costlessly between sectors. Agent j chooses location in order to maximize a discounted sum of payoffs given subjective beliefs $(\alpha_t^j)_{t\geq 0}$ for the time path of shares, where to be consistent we require that $\alpha_0^j = \alpha_0$ for all j.

Clearly, if $\alpha_0 < \alpha_b$, then switching to or remaining in the traditional sector at the end of time zero is a dominant strategy regardless of beliefs, because $r_1 = f(\alpha_0) < f(\alpha_b) = 0$. The collective result of these individual decisions is that $\alpha_1 = 0$. But then $\alpha_1 < \alpha_b$, and the whole process repeats. Thus $\alpha_t = 0$ for all $t \geq 1$. This outcome is interesting, because even the most optimistic set of beliefs lead to the low equilibrium when $f(\alpha_0) < 0$. To the extent that Adsera and Ray's analysis is correct, history must always determine outcomes. [42]

Another way that history can re-enter the equation is if we admit some deviation from perfect rationality and perfect information. As was stressed in the introduction, this takes us back to the role of institutions, through which history is transmitted to the present.

It is reasonable to entertain such deviations here for a number of reasons. First and foremost, assumptions of complete information and perfect rationality are usually justified on the basis of experience. Rationality obtains by repeated observation, and by the punishment of deviant behavior through the carrot and stick of economic payoff. Rational expectations are justified by appealing to *induction*. Agents are assumed to have had many observations from a stationary environment. Laws of motion and hence conditional expectations are inferred on the basis of repeated transition sampling *from every relevant state/action pair* (Lucas 1986). When attempting to break free from a poverty trap, however, agents have most likely *never* observed a transition to the high level equilibrium. On the basis of what experience are they to assess its like-

[41] See, for example, Krugman (1991) or Matsuyama (1991).

[42] There are a number of possible criticisms of the result, most of which are discussed in detail by the authors. If, for example, there are congestion costs or first mover advantages, then moving immediately to the modern sector might be rational for some optimistic beliefs and specification of parameters.

lihood from each state and action? How will they assess the different costs or benefits?

In a boundedly rational environment with limited information, outcomes will be driven by norms, institutions and conventions. It is likely that *these* factors are among the most important in terms of a society's potential for successful coordination on good equilibria. In fact for some models we discuss below the equilibrium choice is not between traditional technology and the modern sector, but rather is a choice between predation (corruption) and production, or between maintaining kinship bonds and breaking them. In some sense these choices are inseparable from the social norms and institutions of the societies within which they are framed. [43]

The central role of institutions may not prevent rapid, successful coordination on good equilibria. After all, institutions and conventions are precisely how societies solve coordination problems. As was emphasized in the introduction, however, norms, institutions and conventions are path dependent by definition. And, in the words of Matsuyama (1995, p. 724), "coordinating expectations is much easier than coordinating changes in expectations." Because of this, economies that start out in bad equilibria may find it difficult to break free.

Why should a convention that locks an economy into a bad equilibrium develop in the first place? Perhaps this is just the role of historical accident. Or perhaps, as Sugden (1989) claims, conventions tend to spread on the basis of versatility or analogy. [44] If so, the conventions that propagate themselves most successfully may be those which are most versatile or susceptible to analogy—not necessarily those which lead to "good" or efficient equilibria.

Often the debate on historical conditions and coordination is cast as "history versus expectations." We have emphasized the role of history, channeled through social norms and institutions, but without intending to say that beliefs are not important. Rather, beliefs are no doubt crucial. At the same time,

[43] More traditional candidates for coordinating roles among the set of institutions include interventionist states promoting industrialization through policy-based financing, or (the cultural component of) large business groups, such as Japan's keiretsu and South Korea's chaebol. In Section 5.2, we discuss the potential for large banks with significant market power to drive "big push" type investments by the private sector.

[44] A versatile convention works reasonably well against many strategies, and hence is advantageous when facing great uncertainty. Analogy means that a rule for a particular situation is suggested by similar rules applied to different but related situations.

beliefs and expectations are shaped by history. And they in turn combine with value systems and local experience to shape norms and institutions. The latter then determine how successful different societies are in solving the particular coordination problems posed by interactions in free markets.

If beliefs and expectations are shaped by history, then the "history versus expectations" dichotomy is misleading. The argument that beliefs and expectations are indeed formed by a whole variety of historical experiences has been made by many development theorists. In an experiment investigating the effects of the Indian caste system, Hoff and Pandey (2004) present evidence that individuals view the world through their own lens of "historically created social identities," which in turn has a pronounced effect on expectations. Rostow (1990, p. 5) writes that "the value system of [traditional] societies was generally geared to what might be called a long run fatalism; that is, the assumption that the range of possibilities open to one's grandchildren would be just about what it had been for one's grandparents." Ray (2003, p. 1) argues that "poverty and the failure of aspirations may be reciprocally linked in a self-sustaining trap."

Finally, experimental evidence on coordination games with multiple Pareto-ranked equilibria suggests that history is important: Outcomes are strongly path dependent. For example, Van Huyck, Cook and Battalio (1997) study people's adaptive behavior in a generic game of this type, where multiple equilibria are generated by strategic complementarities. In each experiment, eight subjects participated in a sequence of between 15 and 40 plays. The authors find sensitivity to initial conditions, defined here as the median of the first round play. In their view, "the experiment provides some striking examples of coordination failure growing from small historical accidents."

4 Empirics of Poverty Traps

Casual observation of the cross-country income panel tends to be suggestive of mechanisms which reinforce wealth or poverty. In Section 4.1 we review the main facts. Section 4.2 considers tests for the empirical relevance of poverty trap models. While the results of the tests support the hypothesis that the map from fundamentals to economic outcomes is not unique, it gives no indication as to what forces might be driving multiplicity. Section 4.3 begins the difficult task of addressing this issue in a macroeconomic framework. Finally, Section 4.4 gives references to empirical tests of specific microeconomic

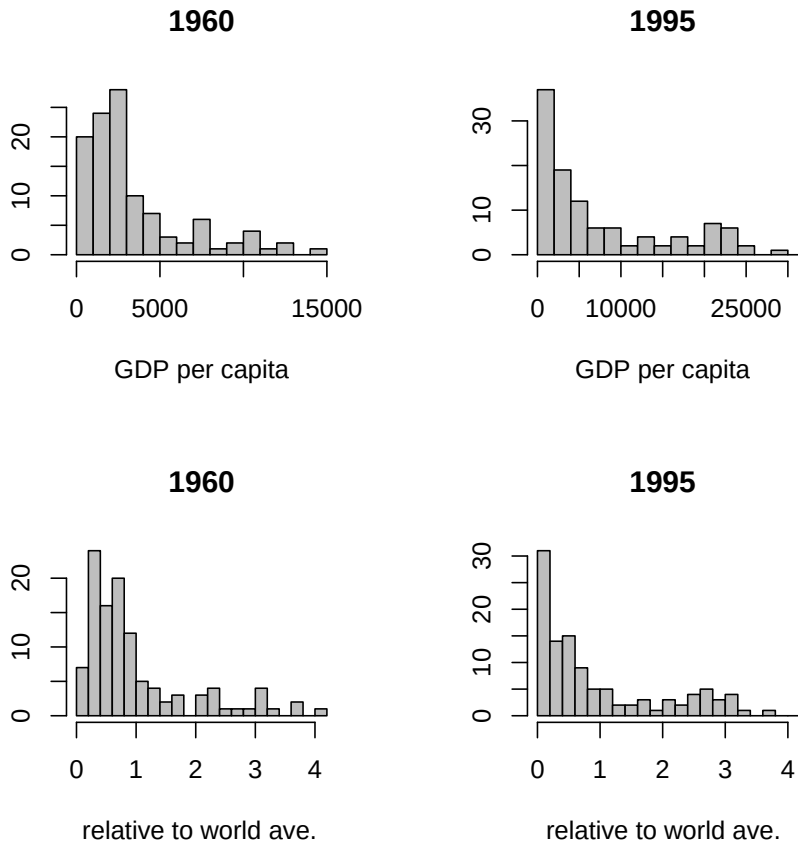


Fig. 14. Absolute and relative GDP per capita

mechanisms that can reinforce poverty at the individual or group level.

4.1 Bimodality and convergence clubs

A picture of the evolving cross-country income distribution is presented in Figure 14. For both the top and bottom histograms the y-axis measures frequency. For the top pair (1960 and 1995) the x-axis is GDP per capita in 1996 PPP adjusted dollars. This is the standard histogram of the cross-country income distribution. For the bottom pair the x-axis represents income as a fraction of the world average for that year.

The single most striking feature of the absolute income histograms for 1960 and 1995 is that over this period a substantial fraction of poor countries have grown very little or not at all. At the same time, a number of middle income countries have grown rapidly, in some cases fast enough to close in on the rich. Together, these forces have caused the distribution to become somewhat

thinner in the middle, with probability mass collecting at the extremes. Such an outcome is consistent with mechanisms that accentuate differences in initial conditions, and reinforce wealth or poverty. Related empirical studies include Azariadis and Drazen (1990), Quah (1993, 1996), Durlauf and Johnson (1995), Bianchi (1997), Pritchett (1997), Desdoigts (1999) and Easterly and Levine (2000).

As well as observing past and present distributions, Quah (1993) also used the Penn World Tables to estimate a transition probability matrix by discretizing the state space (income per capita), treating all countries as observations from the same Markovian probability law, and measuring transition frequency. This matrix provides information on mobility. Also, by studying the ergodic distribution, and by multiplying iterations of the matrix with the current cross-country income distribution, some degree of inference can be made as to where the income distribution is heading.

In his calculation, Quah uses per capita GDP relative to the world average over the period 1962 to 1984 in a sample of 118 countries. Relative income is discretized into state space $S := \{1, 2, 3, 4, 5\}$ consisting of 5 "bins," with states corresponding to values for relative GDP of 0–0.25, 0.25–0.5, 0.5–1, 1–2 and 2–∞ respectively. The transition matrix $\mathbf{P} = (p_{ij})$ is computed by setting p_{ij} equal to the fraction of times that a country, finding itself in state i, makes the transition to state j the next year. The model is assumed to be stationary, so all of the transitions can be pooled when calculating transition probabilities. The result of this calculation (Quah 1993, p. 431) is

$$\mathbf{P} = \begin{pmatrix} 0.97 & 0.03 & 0.00 & 0.00 & 0.00 \\ 0.05 & 0.92 & 0.03 & 0.00 & 0.00 \\ 0.00 & 0.04 & 0.92 & 0.04 & 0.00 \\ 0.00 & 0.00 & 0.04 & 0.94 & 0.02 \\ 0.00 & 0.00 & 0.00 & 0.01 & 0.99 \end{pmatrix}.$$

The Markov chain represented by $\mathbf{P}$ is easily shown to be ergodic, in the sense that there is a unique $\psi^* \in \mathscr{P}(S)$, the distributions on S, with the property that $\psi^*\mathbf{P} = \psi^*$, and $\psi\mathbf{P}^t \to \psi^*$ as $t \to \infty$ for all $\psi \in \mathscr{P}(S)$.[45] Quah calculates this ergodic distribution ψ^* to be $(0.24, 0.18, 0.16, 0.16, 0.27)$.

[45] Following Markov chain convention we are treating the distributions in $\mathscr{P}(S)$ as row vectors. Also, $\mathbf{P}^t$ is t compositions of $\mathbf{P}$ with itself. For more discussion of ergodicity see the technical appendix, or Stachurski (2004).

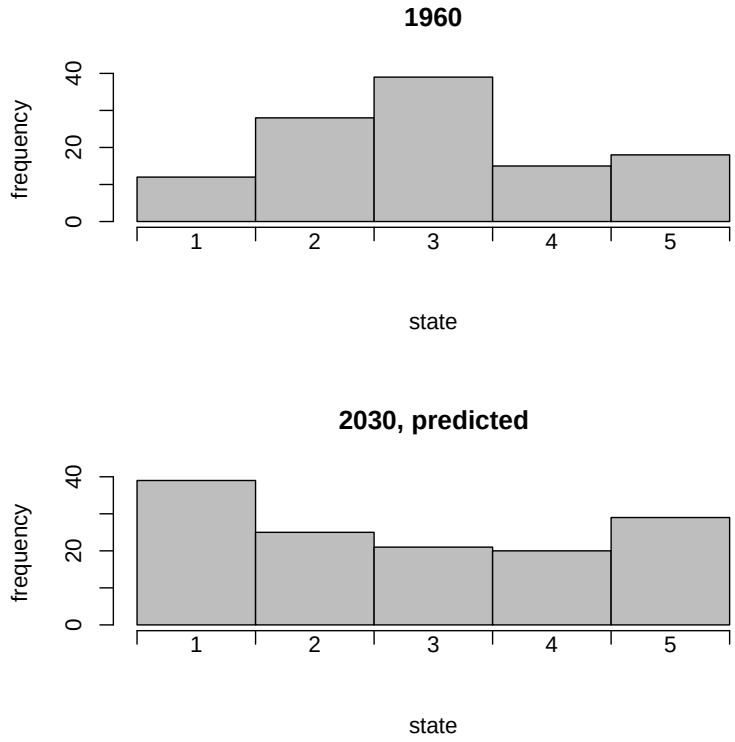


Fig. 15. Discrete projection

The ergodic distribution is quite striking, in that the world is divided almost symmetrically into two convergence clubs of rich and poor at either end of the income distribution.

It is not immediately clear just how long the long run is. To get some indication, we can apply $\mathbf{P}^t$ to the current distribution for different values of t. Figure 15 shows the results of applying $\mathbf{P}^{30}$ to the year 2000 income distribution from the Penn World Tables. This gives a projection for the 2030 distribution. Contrasted with the 1960 distribution the prediction is strongly bimodal.

As Quah himself was at pains to emphasize, the projections carried out above are only a first pass at income distribution dynamics, with many obvious problems. One of those is that the dynamics generated by a discretized version of a continuous state Markov chain can deviate very significantly from the true dynamics generated by the original chain, and error bounds are difficult to quantify. [46] Also, since the estimation of $\mathbf{P}$ is purely nonparametric, the projections do not contain any of the restrictions implied by growth theory.

[46] Compare, for example, Feyrer (2003) and Johnson (2004).

Quah (1996) addressed the first of these problems by estimating a continuous state version. In the language of this survey, he estimates a stochastic kernel Γ, of which $\mathbf{P}$ is a discretized representation. The estimation is nonparametric, using a Parzen-window type density smoothing technique. The kernel is suggestive of considerable persistence.

Azariadis and Stachurski (2004) make some effort to address both the discretization problem and the lack of economic theory simultaneously, by estimating Γ *parametrically*, using a theoretical growth model. In essence, they estimate equation (9), where $k \mapsto A(k)$ is represented by a three-parameter logistic function. The logistic function nests a range of growth models, from the convex model in Figure 2 to the nonconvex models in Figure 7, panels (a), (b) and (d). Once the law of motion (9) is estimated, the stochastic kernel Γ is calculated via equation (10), and the projection of distributions is computed by iterating (4).

The resulting 2030 prediction is shown in Figure 16, with the 1960 distribution drawn above for comparison. The x-axis is log of real GDP per capita in 1996 US dollars. The 1960 density is just a smoothed density estimate using Gaussian kernels, with data from the Penn World Tables. The same data was used to estimate the parameters in the law of motion (9). As in Figure 15, a unimodal distribution gives way to a bimodal distribution.

These findings do lend some support to Quah's convergence club hypothesis. Much work remains to be done. For example, in all of the methodologies discussed above, nonstationary data is being fitted to a stationary Markov chain. This is clearly a source of bias. Furthermore, all of these models are too small, in the sense that the state space used in the predictions are only one-dimensional. [47]

[47] In fact within each economy there are many interacting endogenous variables, only one of which is income. Even if the process as a whole is stationary and Markov, projection of the system onto one dimension will yield a process which is not generally Markovian. Moreover, there are interactions between countries that affect economic performance, and these interactions are important. A first-best approach would be to treat the world economy as an $N \times M$-dimensional Markov process, where N is the number of countries, and M is the number of endogenous variables in each country. One would then estimate the stochastic kernel Γ for this process, a map from $\mathbb{R}_+^{N \times M} \times \mathbb{R}_+^{N \times M} \to [0, \infty)$. Implications for the cross-country income distribution could be calculated by computing marginals.

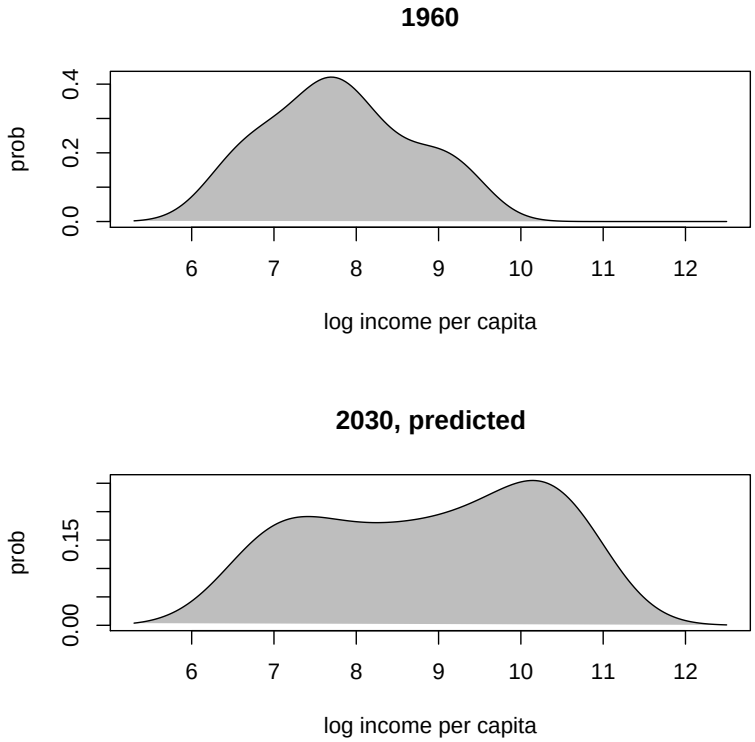


Fig. 16. Parametric projection

4.2 Testing for existence

Poverty trap models tend to be lacking in testable quantitative implications. Where there are multiple equilibria and sensitive dependence to initial conditions, outcomes are much harder to pin down than when the map from parameters to outcomes is robust and unique. This has led many economists to question the empirical significance of poverty trap models. [48] In this section, we ask whether or not there is any evidence that poverty traps exist.

In answering this question, one must be very careful to avoid the following circular logic: First, persistent poverty is observed. Poverty traps are then offered as the explanation. But how do we know there are poverty traps? Because (can't you see?) poverty persists. [49] This simple point needs to be kept in mind when interpreting the data with a view to assessing the empirical relevance of the models in this survey. Persistent poverty, emergent bimodality and the dispersion of cross-country income are the phenomena we seek to explain. They cannot themselves be used as proof that poverty traps explain

[48] See Matsuyama (1997) for more discussion of this point.
[49] This is a version of Karl Popper's famous tale about Neptune and the sea.

the data.

Also, a generalized convex neoclassical model can certainly be the source of bimodality and dispersion if we accept that the large differences in total factor productivity residuals across countries are due to some exogenous force, the precise nature of which is still waiting to be explained. In this competing explanation, the map from fundamentals to outcomes is unique, and shocks or historical accidents which perturb the endogenous variables can safely be ignored.

The central question, then, is *whether or not the poverty trap explanation of cross-country income differentials survives if we control for the exogenous forces which determine long run economic performance.* In other words, do self-reinforcing and path dependent mechanisms imply that economies populated by fundamentally similar people in fundamentally similar environments can support very different long run outcomes? What empirical support is there for such a hypothesis?

One particularly interesting study which addresses this question is that of Bloom, Canning and Sevilla (2003). Their test is worth discussing in some detail. To begin, consider again the two multiple equilibria models shown in Figure 8 (page 28), along with their ergodic distributions. As can be seen in the left hand panels, when the shock is suppressed both Country A and Country B have two locally stable equilibria for capital per worker—and therefore two locally stable equilibria for income. Call these two states y_1^* and y_2^*, the first of which is interpreted as the poverty trap.

In general, y_1^* and y_2^* will depend on the vector of exogenous fundamentals, which determine the exact functional relationships in the model, and hence become parameters in the law of motion. Let this vector be denoted by $\mathbf{x}$. Consider a snapshot of the economy at some point in time t. We can write income per capita as

$$y = \begin{cases} y_1^*(\mathbf{x}) + u_1 & \text{with probability } p(\mathbf{x}); \\ y_2^*(\mathbf{x}) + u_2 & \text{with probability } 1 - p(\mathbf{x}). \end{cases} \tag{R2}$$

Here $p(\mathbf{x})$ is the probability that the country in question is in the basin of attraction for the lower equilibrium $y_1^*(\mathbf{x})$ at time t. This probability is determined by the time t marginal distribution of income. The shock u_i represents deviation from the deterministic attractor at time t.

Figure 8 (page 28) helps to illustrate how y_1^* and y_2^* might depend on the

exogenous variables. Imagine that Countries A and B have characteristics $\mathbf{x}_A$ and $\mathbf{x}_B$ respectively. These different characteristics account for the different shapes of the laws of motion shown in the left hand side of the figure. As drawn, $y_2^*(\mathbf{x}_A)$, the high level attractor for Country A, is less than $y_2^*(\mathbf{x}_B)$, the high level attractor for Country B, while $y_1^*(\mathbf{x}_A)$ and $y_1^*(\mathbf{x}_B)$ are roughly equal.

In addition, we can see how the probability $p(\mathbf{x})$ of being in the poverty trap basin depends on these characteristics. For time t sufficiently large, ergodicity means that the time t marginal distribution—which determines this probability—can be identified with the ergodic distribution. The ergodic distribution in turn depends on the underlying structure, which depends on $\mathbf{x}$. This is illustrated by the different sizes of the distribution modes for Countries A and B in Figure 8. For Country A the left hand mode is relatively large, and hence so is $p(\mathbf{x})$.

Using a maximum likelihood ratio test, the specification (R2) is evaluated against a single regime alternative

$$y = y^*(\mathbf{x}) + u, \tag{R1}$$

which can be thought of for the moment as being generated by a convex Solow model. The great benefit of the specification (R1) and (R2)—as emphasized by the authors—is that long run output depends only on exogenous factors. The need to specify the precise system of endogenous variables and their interactions is circumvented. [50]

In conducting the test of (R1) against (R2), it is important *not* to include as exogenous characteristics any variable which is in fact endogenously determined. For to do so might result in conditioning on the *outcomes* of the underlying process which generates multiple equilibria. In the words of the authors, "Including such variables may give the impression of a unique equilibrium relationship [for the economic system] when in reality they are a function of the equilibrium being observed. Fundamental forces must be characteristics that determine a country's economic performance but are not determined by it."

In the estimation of Bloom, Canning and Sevilla, only geographic features are included in the set of exogenous variables. These include data on distance from equator, rainfall, temperature, and percentage of land area more than 100km from the sea. For this set of variables, the likelihood ratio test rejects the single regime model (R1) in favor of the multiple equilibria model (R2).

[50] Ergodicity is critical in this respect, for without it p will depend not just on $\mathbf{x}$ but also on the lagged values of endogenous variables.

They find evidence for a high level equilibrium which does not vary with $\mathbf{x}$, and a low level equilibrium which does. In particular, $y_1^*(\mathbf{x})$ tends to be smaller for hot, dry, land-locked countries (and larger for those with more favorable geographical features). In addition, $p(\mathbf{x})$ is larger for countries with unfavorable geographical features. In other words, the mode of the ergodic distribution around $y_1^*(\mathbf{x})$ is relatively large. For these economies escape from the poverty trap is more difficult.

Overall, the results of the study support the poverty trap hypothesis. They also serve to illustrate the importance of distinguishing between variables which are exogenous and those which have feedback from the system. If one conditions on "explanatory" variables which deviate significantly from fundamental forces, the likelihood of observing multiple equilibria in the map from those variables to outcomes will be lower. For example, one theme of this survey is that institutions can be an important source of multiplicity, either directly or indirectly through their interactions with the market. If institutions are endogenous, and if traps in institutions drive the disparities in cross-country incomes, then conditioning on institutions may give spurious convergence results entirely disconnected from long run outcomes generated by the system.

4.3 Model calibration

One of the advantages of the methodology proposed by Bloom, Canning and Sevilla is that estimation and testing can proceed without fully specifying the underlying model. The exacting task of determining the relevant set of endogenous variables and the laws by which they interact is thereby circumvented. But there are two sides to this coin. While the results of the test suggest that poverty traps matter, they give no indication as to their source, or to the appropriate framework for formulating them as models.

Graham and Temple (2004) take the opposite approach. They give the results of a numerical experiment starting from a specific poverty trap model, somewhat akin to the inertial self-reinforcement model of Section 3.4. The question they ask is whether or not the model in question has the potential to explain observed cross-country variation in per capita income for a reasonable set of parameters. We briefly outline their main findings, as well as their technique for calibration, which is of independent interest.

As in Section 3.4, there is both a traditional agricultural sector and a modern sector with increasing social returns due to technical externalities. The

agricultural sector has a decreasing returns technology

$$Y_a = A_a L_a^{\gamma}, \quad \gamma \in (0,1), \tag{11}$$

where Y_a is output, A_a is a productivity parameter and L_a is labor employed in the agricultural sector. The j-th firm in the modern sector has technology

$$Y_{m,j} = A_m L_{m,j} L_m^{\lambda}, \quad \lambda > 0, \tag{12}$$

where $Y_{m,j}$ is output of firm j, A_m is productivity, $L_{m,j}$ is labor employed by firm j, and L_m is total employment in the modern sector. The firm ignores the effect of its hiring decisions on L_m, thus setting the stage for multiplicity. We set $L_a + L_m = L$, a fixed constant, and, as usual, $\alpha := L_m/L$.

The relative price of the two goods is fixed in world markets and normalized to one by appropriate choice of units. Wages are determined by marginal cost pricing: $w_a = \gamma A_a L_a^{\gamma-1}$ and $w_m = A_m L_m^{\lambda}$. Setting these factor payments equal gives the set of equilibrium modern sector shares α as solutions to the equation

$$(1-\alpha)^{1-\gamma}\alpha^{\lambda} = \frac{A_a \gamma L^{\gamma-1-\lambda}}{A_m}. \tag{13}$$

Regarding calibration, γ is a factor share, and the increasing returns parameter λ has been calculated in several econometric studies.[51] Relative productivity is potentially more problematic. However, it turns out that (13) has precisely two solutions for reasonable parametric values. Since both solutions α_1 and α_2 satisfy (13) we have

$$(1-\alpha_1)^{1-\gamma}\alpha_1^{\lambda} - (1-\alpha_2)^{1-\gamma}\alpha_2^{\lambda} = 0. \tag{14}$$

In which case, assuming that current observations are in equilibrium, one can take the observed share as α_1, calculate α_2 as the other solution to (14), and set the poverty trap equilibrium equal to $\alpha_1^* := \min\{\alpha_1, \alpha_2\}$. The high productivity equilibrium is $\alpha_2^* := \max\{\alpha_1, \alpha_2\}$. Figure 17 illustrates this procedure for $\alpha_1 = 0.1$, $\gamma = 0.7$ and $\lambda = 0.3$.

When α_1^*, α_2^*, γ and λ are known, a little algebra shows that the ratio of output in the high equilibrium to output in the low equilibrium can also be computed. In this way it is possible to evaluate the relative impact of the poverty trap on individual countries and the cross-country income distribution.

Using this strategy and a more elaborate model (including both capital and land), Graham and Temple's main findings are as follows. First, for reasonable

[51] See, for example, Cabarello and Lyons (1992).

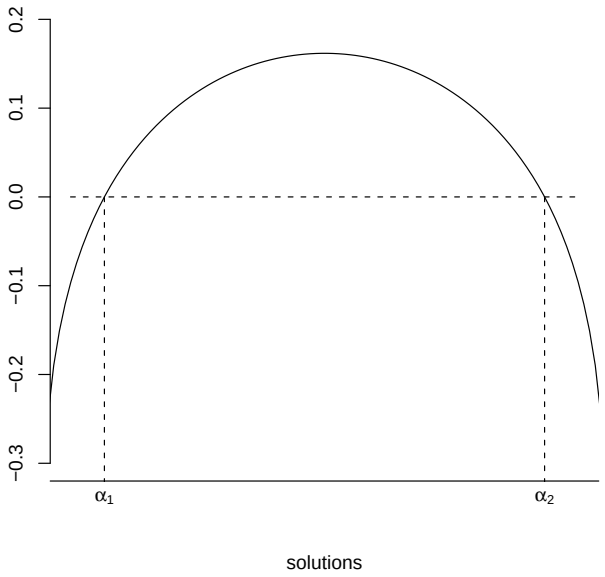


Fig. 17. Multiple equilibria

parameter values some 1/4 of the 127 countries in their 1988 data set are in the poverty trap α_1^*. Second, after calculating the variance of log income across countries when all are in their high output equilibrium and comparing it to the actual variance of log income, they find that the poverty trap model can account for some 2/5 to one half of all observed variation in incomes.

Overall, their study suggests that the model can explain some properties of the data, such as the difference between poor, agrarian economies and low to middle income countries. On the other hand, it cannot account for the huge differences between the very poorest and the rich industrialized countries. In the model, the largest ratios of low to high equilibrium production are in the region of two to three. As we saw in Section 2.1, however, actual per capita output ratios between rich and poor countries are much larger.

4.4 Microeconomic data

There has also been research in recent years on poverty traps that occur at the individual or group level. For example, Jalan and Ravallion (2002) fit a microeconomic model of consumption growth with localized spillovers from capital to farm-household panel data in rural China. Their results are consistent with empirical significance of geographical poverty traps. Other authors

have studied particular trap mechanisms. For example, Bandiera and Rasul (2003) and Conley and Udry (2003) consider the effects of positive network externalities on technology adoption in Mozambique and Ghana respectively. Barrett, Bezuneh and Aboud (2001) consider the dynamic impact of credit constraints on the poor in Côte d'Ivoire and Kenya. Morduch (1990) studies the effect of risk on income in India, as does Dercon (1998) for Tanzania.

5 Nonconvexities, Complementarities and Imperfect Competition

Increasing returns production under imperfect competition is a natural framework to think about multiple equilibria. Imperfect competition leads directly to externalities transmitted through the price system, because monopolists themselves, rather than Walrasian auctioneers, set prices, and presumably they do so with their own profit in mind. At the same time, their pricing and production decisions impinge on other agents. These general equilibrium effects can be a source of multiplicity.

Section 5.1 illustrates this idea using the big push model of Murphy, Shleifer and Vishny (1989); a model which formalizes an earlier discussion in Rosenstein-Rodan (1943). Rosenstein-Rodan argued that modern industrial technology is freely available to poor countries, but has not been adopted because the domestic market is too small to justify the fixed costs it requires. If all sectors industrialize simultaneously, however, the market may potentially be expanded to the extent that investment in modern technology is profitable.

Thus the big push model of Section 5.1 helps to clarify the potential challenges posed by coordination for the industrialization process. We shall see that the major coordination problem facing monopolists cannot be resolved by the given market structure. In this situation, the ability of a society to successfully coordinate entrepreneurial activity—and thereby realize the social benefits available in modern production technologies—will depend in general on such structures as its institutions, political organizations, the legal framework, and social and business conventions.

In countries such as South Korea, the state has been very active in attempting to overcome coordination problems associated with industrialization. In Western Europe, the state was typically much less active, and the role of the private sector was correspondingly larger. For example, Da Rin and Hellmann (2002) have recently emphasized the important role played by banks in coordinating industrialization. Section 5.2 reviews their model.

A theme of this survey is traps that prevent economies as a whole from adopting modern production technologies. One aspect of this transformation to modernity is the need for human capital. If investment in human capital has a high economic payoff then a skilled work-force should spontaneously arise. Put differently, if the poor are found to invest little in schooling or training then this suggests to us that returns to these investments are relatively low. Section 5.3 reviews Kremer's (1993) matching model, where low investment in schooling sustains itself in a self-reinforcing trap.

Finally, Section 5.4 gives references to notable omissions on the topic of increasing returns.

5.1 Increasing returns and imperfect competition

Murphy, Shleifer and Vishny's (1989) formalization of Rosenstein-Rodan's (1943) big push is something of a watershed in development economics. Their model turns on demand spillovers which create complementarities to investment. They point out that for the economy to generate multiple equilibria, it must be the case that investment simultaneously (i) increases the size of other firms' markets, or otherwise improves the profitability of investment; and (ii) has negative net present value. This means that profits alone cannot be the direct source of the market size effects; otherwise (i) and (ii) would be contradictory.

In the first model they present, higher wages in the modern sector are the channel through which demand spillovers increase market size. Although investment is not individually profitable, it raises labor income, which in turn raises the demand for other products. If the spillovers are large enough, multiple equilibria will occur. In their second model, investment in the modern technology changes the composition of aggregate demand across time. In the first period, the single monopolistic firm in each sector decides whether to invest or not. Doing so incurs a fixed cost F in the first period, and yields output ωL in the second, where $\omega > 1$ is a parameter and L is labor input. The cost in the second period is just L, as wages are the numeraire. If, on the other hand, the monopolist chooses not to invest, production in that sector will take place in a "competitive fringe" of atomistic firms using constant returns to scale technology. For these firms, one unit of labor input yields one unit of output. The price for each unit so produced is unity.

All wages and profits accrue to a representative consumer, who supplies L

units of labor in both periods, and maximizes the undiscounted utility of his consumption, that is,

$$\max\left\{\int_0^1 \ln c_1(\alpha)d\alpha + \int_0^1 \ln c_2(\alpha)d\alpha \;\Big|\; c_t\colon [0,1] \ni \alpha \mapsto c_t(\alpha) \in [0,\infty)\right\}$$

subject to the constraints $\int_0^1 c_1 p_1 \leq y_1$ and $\int_0^1 p_2 c_2 \leq y_2$. Here $\alpha \in [0,1]$ indexes the sector, $c_t(\alpha)$ and $p_t(\alpha)$ are consumption and price of good α at time t respectively, and y_t is income (wages plus profits) at time t. [52]

In the first period only the competitive fringe produces, and $p_1(\alpha) = 1$ for all α. In the second, monopolists face unit elastic demand curves $c_2(\alpha) = y_2/p_2(\alpha)$. Given these curves and the constraints imposed by the competitive fringe, monopolists set $p_2(\alpha) = 1$ for all α. Their profits are $\pi = ay_2 - F$, where $a := 1 - 1/\omega$ is the mark-up.

Consider profitability when all entrepreneurs corresponding to sectors $[0,\alpha]$ decide to invest. (The number α can also be thought of as the fraction of the total number of monopolists who invest.) It turns out that for some parameter values both $\alpha = 0$ and $\alpha = 1$ are equilibria. To see this, consider first the case $\alpha = 0$, so that $y_1 = y_2 = L$. It is not profitable for a firm acting alone to invest if $\pi = aL - F \leq 0$. On the other hand, if $\alpha = 1$, then $y_1 = L - F$ and $y_2 = \omega L$, so monopolists make positive profits when $a\omega L - F \geq 0$. Multiple equilibria exist if these inequalities hold simultaneously. In Figure 18 multiple equilibria obtain for all $L \in [L_1, L_2]$.

As was mentioned in the introduction, coordination problems and other mechanisms that reinforce the status quo can interact with each other and magnify their individual impact. Murphy, Shleifer and Vishny (1989) provide a simple example of this in the context of the model outlined above. They point out that the coordination problem for the monopolists is compounded if industrialization requires widespread development of infrastructure and intermediate inputs, such as railways, road networks, port facilities and electricity grids. All of these projects will themselves need to be coordinated with industrialization.

For example, suppose that n infrastructure projects must be undertaken in the first period to permit industrialization in the second. Each project has a fixed

[52] To simplify the exposition we assume that consumers can neither save nor dissave from current income. For the moment we also abstract from the existence of a financial sector. Firms which invest simply pay all wages in the second period at a zero rate of interest. See the original for a more explicitly general equilibrium formulation.

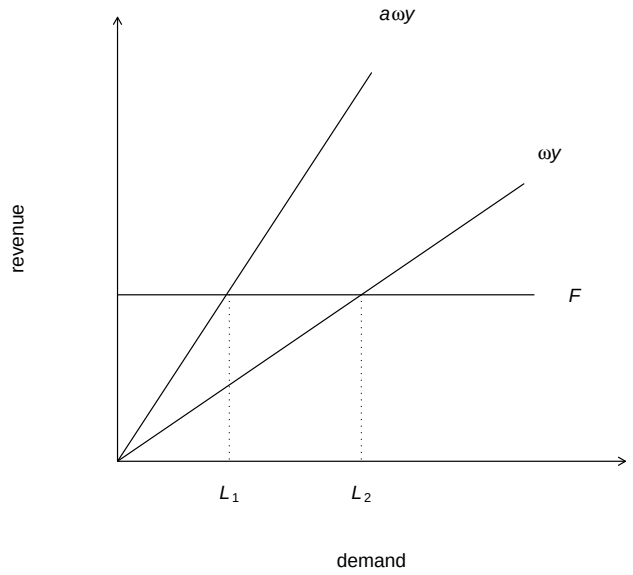


Fig. 18. Multiple equilibria

cost R_n, and operates in the second period at zero marginal cost. Leaving aside the issue of how the spoils of industrialization will be divided among the owners of the projects and the continuum of monopolists, it is clear that industrialization has the potential to be profitable for all only when $a\omega L - F$, the profits of the monopolists when $\alpha = 1$, exceed total infrastructure costs $\sum_{i=1}^{n} R_i$.

If the condition $aL - F \leq 0$ continues to hold, however, individual monopolists investing alone will be certain to lose money. Realizing this, investors in infrastructure face extrinsic uncertainty as to whether or not industrialization will actually take place. Given their subjective evaluation, they may choose not to start their infrastructure projects. In turn, the monopolists are aware that investors in infrastructure face uncertainty, and may themselves refrain from starting projects. This makes monopolists even more uncertain as to whether or not the conditions for successful industrialization will eventuate. The fixed point of this infinite regression of beliefs may well be inaction. In either case, the addition of more actors adds to the difficulty of achieving coordination.

As Da Rin and Hellmann (2002) have recently emphasized, one candidate within the private sector for successfully coordinating a big push type industrialization is the banks. Banks are the source of entrepreneurs' funds, and shape the terms and conditions under which capital may be raised. In addition, banks interact directly with many entrepreneurs. Finally, banks can potentially profit from coordinating industrialization if their market power is large.

Da Rin and Hellmann find that the structure and legal framework of the banking sector are important determinants of its ability to coordinate successful industrialization. To illustrate their ideas, consider again the big push model of Section 5.1. In order to make matters a little easier, let us simply define the second period return of monopolists (entrepreneurs) to be $f(\alpha)$, where α is the fraction of entrepreneurs who decide to set up firms and the function $f\colon [0,1] \to \mathbb{R}$ is strictly increasing. As before, there is a fixed cost F to be paid in the first period, which we set equal to 1. The future is not discounted.

It is convenient to think of the number of entrepreneurs as some large but finite number N.[53] In addition to these N entrepreneurs, there is now a financial sector, members of whom are referred to as either "banks" or "investors." There are $B \in \mathbb{N}$ banks, the first $B-1$ of which have an intermediation cost of r per unit of investment. The last bank has an intermediation cost of zero, but can lend to only $\ell \leq N$ firms. The number ℓ can be thought of as a measure of the last bank's market power.

The equilibrium lending rate at which firms borrow in the first period is determined by the interaction of the monopolists and the banks. In the first stage of the game, each bank b offers a schedule of interest rates to the N firms. This strategy will be written as $\sigma_b := \{i_n^b : 1 \leq n \leq N\}$. The collection of these strategies across banks will be written as $\sigma := \{\sigma_b : 1 \leq b \leq B\}$. Let Σ be the set of all such σ.

In the second stage, each entrepreneur either rejects all offers and does not set up the firm, or selects the minimum interest rate, pays the fixed cost and enters the market. In what follows we write $m_n(\sigma)$ to mean $\min_b i_n^b$, the minimum interest rate offered to firm n in σ. If a fraction α accepts contracts then firm

[53] In particular, entrepreneurs do not take into account their influence on α when evaluating whether to set up firms or not.

n makes profits

$$\pi(\alpha, m_n(\sigma)) = f(\alpha) - (1 + m_n(\sigma)). \tag{15}$$

For bank $b < B$, profits are given by

$$\Pi_b(\sigma_b) = \sum_{n=1}^{N} (i_n^b - r) \mathbb{1}\{\text{firm } n \text{ accepts}\}, \tag{16}$$

where here and elsewhere $\mathbb{1}\{Q\}$ is equal to one when the statement Q is true and zero otherwise. For $b = B$, profits are

$$\Pi_b(\sigma_b) = \sum_{n=1}^{N} i_n^b \mathbb{1}\{\text{firm } n \text{ accepts}\}. \tag{17}$$

In equilibrium, banks never offer interest rates strictly greater than r, because should they do so other banks will always undercut them. As a result, we can and do assume in all of what follows that $m_n(\sigma) \leq r$ for all n. Also, to make matters interesting, we assume that $f(0) < 1 + r < f(1)$, or, equivalently,

$$\pi(0, r) < 0 < \pi(1, r). \tag{18}$$

Firms' actions will depend on their beliefs—in particular, on what fraction α of the N firms they believe will enter. Clearly beliefs will be contingent on the set of contracts offered by banks. Thus a belief for firm n is a map α_n^e from Σ into $[0, 1]$. Given this belief, firm n enters if and only if

$$\pi(\alpha_n^e(\sigma), m_n(\sigma)) \geq 0. \tag{19}$$

Given σ, the set of self-supporting equilibria for the second stage subgame is

$$\Omega(\sigma) := \left\{ \alpha \in [0, 1] \quad : \quad \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{\pi(\alpha, m_n(\sigma)) \geq 0\} = \alpha \right\}. \tag{20}$$

In other words, $\alpha \in \Omega(\sigma)$ if, given the set of offers σ and the belief on the part of all firms that the fraction of firms entering will be α, exactly $\alpha \times 100\%$ of firms find it optimal to enter.

Beliefs are required to be consistent in the sense that $\alpha_n^e(\sigma) \in \Omega(\sigma)$ for all σ and all n. Beliefs are called *optimistic* if $\alpha_n^e = \alpha^{\text{opt}}$ for all n, where $\alpha^{\text{opt}}(\sigma) := \max \Omega(\sigma)$ for all $\sigma \in \Sigma$. In other words, all agents believe that as many firms will enter as are consistent with offer σ, and this is true for every $\sigma \in \Sigma$. Beliefs are defined to be *pessimistic* if the opposite is true; that is, if $\alpha_n^e = \alpha^{\text{pes}}$ for all n, where $\alpha^{\text{pes}}(\sigma) := \min \Omega(\sigma)$ for all $\sigma \in \Sigma$.

Da Rin and Hellmann first observe that if $\ell = 0$, then the outcome of the game will be determined by beliefs. In particular, if beliefs are pessimistic, then the low equilibrium $\alpha = 0$ will obtain. If beliefs are optimistic, then the high equilibrium $\alpha = 1$ will obtain. The interpretation is that when $\ell = 0$, so that the market for financial services is entirely competitive (in the sense of Bertrand competition with identical unit costs described above), the existence of the financial sector will not alter the primary role of beliefs in determining whether industrialization will take place.

Let us verify this observation in the case of pessimistic beliefs. To do so, it is sufficient to show that if $\sigma \in \Sigma$ is optimal, then $0 \in \Omega(\sigma)$. The reason is that if $0 \in \Omega(\sigma)$, then by (20) we have $\pi(0, m_n(\sigma)) < 0$ for all n. Also, beliefs are pessimistic, so $\alpha_n^e(\sigma) = \min \Omega(\sigma) = 0$. In this case no firms enter by (19).

To see that $0 \in \Omega(\sigma)$ for all optimal σ, suppose to the contrary that $\sigma \in \Sigma$ is optimal, but $0 \notin \Omega(\sigma)$. Then $\pi(0, m_k(\sigma)) \geq 0$ for some k, in which case (18) implies that $m_k(\sigma) < r$. Because firms only accept contracts at rates less than r (that is, $m_n(\sigma) \leq r$ for all n), it follows from (16) that the bank which lent to k looses money, and σ is not optimal. The intuition is that no bank has market power, and cannot recoup losses sustained when encouraging firms to enter by offering low interest rates.

More interesting is the case where the last bank B has market power. With sufficient market power, B will induce industrialization (the high equilibrium where $\alpha = 1$) even when beliefs are pessimistic:

Proposition 5.1 (Da Rin and Hellmann) *Suppose beliefs are pessimistic. In this case, there exists an $\bar{\alpha} \in [0, 1]$ depending on r and f such that industrialization will occur whenever ℓ, the market power of B, satisfies $\ell/N \geq \bar{\alpha}$.*

The result shows that rather than relying on spontaneous coordination of beliefs, financial intermediaries may instead be the source of coordination. The key intuition is that a financial intermediary may have a profit motive for inducing industrialization. But to achieve this, two things are necessary: size and market power. Size (as captured by ℓ) is necessary to induce a critical mass of entrepreneurs to invest. Market power (as captured by the cost advantage r) is necessary to recoup the costs of mobilizing that critical mass. We sketch Da Rin and Hellmann's proof in the appendix.

Until now we have considered only the possibility that the banks offer pure debt contracts. Da Rin and Hellmann also study the case where the banks may hold equity as well (i.e., universal banking). They show that in this case

the threshold level at which the lead bank B has sufficient market power to mobilize the critical mass is lower. Industrialization is unambiguously more likely to occur. The reason is that equity permits B to partake in the ex post profits of the critical mass, who benefit from low interest rates on one hand and complete entry ($\alpha = 1$) on the other. With a lower cost of mobilizing firms, B requires less market power to recoup these losses. In Da Rin and Hellmann's words,

> Our model provides a rationale for why a bank may want to hold equity that has nothing to do with the standard reasons of providing incentives for monitoring. Instead, equity allows a bank to participate in the gains that it creates when inducing a higher equilibrium.

In summary, the theory suggests that large universal banks with a high degree of market power can play a central role in the process of industrialization. This theory is consistent with the evidence from countries such as Belgium, Germany and Italy, where a few oligopolist banks with strong market positions played a pivotal role. Some were pioneers of universal banking, and many directly coordinated activity across sectors by participation in management. The theory may also explain why other countries, such as Russia, failed to achieve significant industrialization in the 19th Century. There banks were small and dispersed, their market power severely restricted by the state.

5.3 Matching

The next model we consider is due to Kremer (1993), and has the following features. A production process consists of n distinct tasks, organized within a firm. For our purposes n can be regarded as exogenous. The tasks are undertaken by n different workers, all of whom have their own given skill level $h_i \in [0, 1]$. Here the skill level will be thought of as the probability that the worker performs his or her task successfully. We imagine that if one worker fails in their task the entire process is ruined and output is zero. If all are successful, the outcome of the process is n units of the product. [54] That is,

$$y = n \prod_{i=1}^{n} \mathbb{1}\{\text{worker } i \text{ successful}\}, \quad \mathbb{P}\{\text{worker } i \text{ successful}\} = h_i, \tag{21}$$

[54] Assuming one unit might seem more natural than n, but the latter turns out to be more convenient.

where as before $\mathbb{1}\{Q\} = 1$ if the statement Q is true and zero otherwise. All of the success probabilities are independent, so that $\mathbb{E}(y) = n \prod_i h_i$.

Consider an economy with a unit mass of workers. The distribution of skills across workers is endogenous, and will be discussed at length below. Kremer's first point is that in equilibrium, firms will match workers of equal skill together to perform the process. The intuition is that (i) firms will not wish to pair a work-force of otherwise skilled employees with one relatively unskilled worker, who may ruin the whole process; and (ii) firms with a skilled work-force will be able to bid more for skilled workers, because the marginal value of increasing the last worker's skill is increasing in the skill of the other workers. Thus, for each firm,

$$\mathbb{E}(y) = nh^n, \quad h \text{ the firm's common level of worker skill.} \tag{22}$$

The first thing to notice about this technology is that the expected marginal return to skill is increasing. As a result, small differences in skill can have relatively large effects on output. This may go some way to explaining the extraordinarily large wage differentials between countries. Moreover, for economies with such technology, positive feedback dynamics of the kind considered in Section 3.3 may result, even if the technology for creating human capital is concave.

Another channel for positive feedbacks occurs when matching is imperfect, perhaps because it is costly or the population is finite. Exact matches may not be possible. In that case, there are potentially returns to agglomeration: Skilled people clustering together will decrease the cost of matching, and increase the likelihood of good matches. Also, an initial distribution of skills will tend to persist, because workers will choose skills so as to be where the distribution is thickest. This maximizes their chances of finding good matches. But this is self-reinforcing: Their choices perpetuate the current shape of the distribution.

There is yet another channel that Kremer suggests may lead to multiple equilibrium distributions of skill. This is the situation where skill levels are imperfectly *observed*. We present a simple (and rather extremist) version of Kremer's model. In the first period, workers decide whether to undertake "schooling" or not. This education involves a common cost $c \in (0, 1)$. In the second, firms match workers, produce, and pay out wages. Both goods and labor markets are competitive, and total wages exhaust revenue. Specifically, it is assumed that each worker's wage w is $1/n$-th of firm's output.

Not all of those who undertake schooling become skilled. We assume that the

educated receive a skill level $h = 1$ with probability $p > 1/2$ and $h = 0$ with probability $1 - p$. Those who do not undertake schooling have the skill level $h = 0$. Further, h is not observable, even for workers. Instead, all workers take a test, which indicates their true skill with probability p and the reverse with probability $1 - p$.[55] That is,

$$t := \text{test score} = \begin{cases} h & \text{with probability } p; \\ 1 - h & \text{with probability } 1 - p. \end{cases} \tag{23}$$

Firms then match workers according to the test score t rather than h.

Let $\alpha \in [0, 1]$ denote the fraction of workers who choose to undertake schooling. We will show that for certain values of the parameters p and c, both $\alpha = 0$ and $\alpha = 1$ are equilibria. In doing so, we assume that p is known to all. Also, workers and firms are risk neutral.

Consider first the case where $\alpha = 0$. If the worker undertakes schooling, then, regardless of his skill and test score, his expected wage is $1/n$-th of $n \prod_i h_i$, where his co-workers are drawn from a pool in which the skilled workers have measure zero. That is, $\mathbb{P}\{h_i = 0\} = 1$. It follows that expected output and wage are zero. Since $c > 0$, it is optimal to avoid schooling.[56]

Now consider the agent's problem when $\alpha = 1$. In the second period, the agent will be matched with other workers having the same test score. In either case, computing expected wages is a signal extraction problem. First, using the fact that agents in the pool of potential co-workers have chosen schooling with probability one, the agent can calculate probable skills of a co-worker chosen at random from the population, given their test score:

$$\mathbb{P}\{h = 1 \,|\, t = 1\} = \frac{\mathbb{P}\{h = 1 \text{ and } t = 1\}}{\mathbb{P}\{t = 1\}} = \frac{p^2}{p^2 + (1 - p)^2} =: \theta_p, \tag{24}$$

and,

$$\mathbb{P}\{h = 1 \,|\, t = 0\} = \frac{\mathbb{P}\{h = 1 \text{ and } t = 0\}}{\mathbb{P}\{t = 0\}} = \frac{p(1 - p)}{p(1 - p) + p(1 - p)} = \frac{1}{2}. \tag{25}$$

The worker can use these probabilities to compute expected output and hence wages given the different outcomes of his own test score. In particular, $\mathbb{E}(w \,|\, t =$

[55] We are using the same p as before just to simplify notation.

[56] On the other hand, if skills are perfectly observable, workers who acquire skills will be matched with n workers from the measure zero set of agents having $h = 1$. In that case $w = 1$. Since $c < 1$ it is optimal to choose schooling, and $\alpha = 0$ is not an equilibrium. The same logic works for any $\alpha < 1$

$1) = \theta_p^n$ and $\mathbb{E}(w \,|\, t = 0) = (1/2)^n$. It follows that the expected return to schooling for the agent is

$$\begin{aligned}\mathbb{E}(w \,|\, \text{schooling}) &= \mathbb{E}(w \,|\, t = 0)\mathbb{P}\{t = 0\} + \mathbb{E}(w \,|\, t = 1)\mathbb{P}\{t = 1\} \\ &= \frac{1}{2^n}(1 - p) + \theta_p^n p.\end{aligned}$$

Conversely, $\mathbb{E}(w \,|\, \text{no schooling}) = \frac{1}{2^n}p + \theta_p^n(1 - p)$. Schooling is optimal if

$$\begin{aligned}c < \mathbb{E}(w \,|\, \text{schooling}) - \mathbb{E}(w \,|\, \text{no schooling}) \\ = (2p - 1)(\theta_p^n - (1/2)^n) := c^*(p).\end{aligned}$$

It is easy to see that $c^*(p) > 0$ whenever $p > 1/2$, which is true by assumption. As a result, schooling will be optimal for some sufficiently small c, and $\alpha = 1$ is an equilibrium too. [57]

What are the sources of multiple equilibria in the model? The first is pecuniary externalities in the labor market: When more agents become educated, the probability that the marginal worker can successful match with a skilled co-worker increases. In turn, this increases the returns to education. [58] Second, there is imperfect information: Skilled workers cannot readily match with other skilled workers. Instead, matching is probabilistic, and depends on the overall distribution of skills. Finally, the increasing expected marginal reward for skill inherent in the production function means that the wage spillovers from the decisions of other agents are potentially large.

Another important model of human capital investment with multiple equilibria is Acemoglu (1997). He shows how labor market frictions can induce a situation where technology adoption is restricted by a lack of appropriately skilled workers. Low adoption in turn reduces the expected return to training, further exacerbating the scarcity of workers who are trained. In other words, poor technology adoption and low capital investment are self-reinforcing, because they cause the very shortage of skilled workers necessary to make such investments profitable.

[57] It may seem that if $p = 1$ and observation is perfect, then $\mathbb{E}(w \,|\, \text{schooling}) - \mathbb{E}(w \,|\, \text{no schooling})$ should be zero, so that no multiple equilibria are possible. But under this assumption the above derivation of $c^*(p)$ is not valid, because we would be conditioning on sets with probability zero.

[58] In fact the expected wage is increased for all, but those who become skilled benefit more.

Young's (1928) famous paper on increasing returns notes that not only does the degree of specialization depend on the size of the market, but the size of the market also depends on the degree of specialization. In other words, there are efficiency gains from greater division of labor, primarily due to application of machines. Greater specialization increases productivity, which then expands the market, leading back into more specialization, and so on. As a result, there are complementarities in investment. These complementarities can be the source of poverty traps. A detailed discussion of this process is omitted from the present survey, but only because excellent surveys already exist. See in particular Matsuyama (1995) and Matsuyama (1997). Other references include Matsuyama and Ciccone (1996), Rodríguez-Clare (1996) and Rodrik (1996).

Increasing returns are also associated with geographical agglomeration. Starrett (1978) points out that agglomerations *cannot* form as the equilibria of perfectly competitive economies set in a homogeneous space. Thus all agglomerations must be caused either by exogenous geographical features or by some market imperfection. An obvious candidate is increasing returns. (It is difficult to see what geographical features could explain the extent of concentration witnessed in places such as Tokyo or Hong Kong.) This survey does not treat geography and its possible connections with poverty traps in much detail. Interested readers might start with the review of Ottaviano and Thisse (2004). [59]

Another source of complementarities partly related to geography is positive network externalities in technology adoption. These are often thought to arise from social learning: Local experience with a technology allows the cost of adoption to decrease as the number of adopters in some network gets larger. As well as information spillovers, more adopters of a given technology may lead to the growth of local supply networks for intermediate inputs, repairs and servicing, skilled labor and so on. See, for example, Beath, Katsoulacos and Ulph (1995), Bandiera and Rasul (2003), Conley and Udry (2003), and Baker (2004).

Finally, an area that we have not treated substantially in this survey is *optimal* growth under nonconvexities, as opposed to the fixed savings rate model

[59] See also Limao and Venables (2001) or Redding and Venables (2004) for the empirics of geography and international income variation.

considered in Section 3.3. In other words, how do economies evolve when (i) agents choose investment optimally by dynamic programming, given a set of intertemporal preferences; and (ii) the aggregate production function is nonconvex?

There are two main cases. One is that increasing returns are taken to be external, perhaps as a feedback from aggregate capital stock to the productivity residual, and agents perceive the aggregate production function to be *convex*. In this case there is a subtle issue: In order to optimize, agents must have a belief about how the productivity residual evolves. This may or may not coincide with its actual evolution as a result of their choices. An equilibrium transition rule is a specification of savings and investment behavior such that (a) agents choose this rule given their beliefs; and (b) those choices cause aggregate outcomes to meet their expectations. Existence of such an equilibrium is far from assured. See Mirman, Morand and Reffett (2004) and references therein. Dynamics are still actively being investigated.

The second case is where increasing returns are internal, and agents perceive aggregate production possibilities exactly as they are. These models generate similar poverty traps as were found for fixed savings rates in Section 3.3. The literature is large. An early investigation is Skiba (1978). See also Dechert and Nishimura (1983), who consider a per capita production function $k \mapsto f(k)$ which is convex over a lower region of the state space (capital per worker), and concave over the remainder; and Amir, Mirman and Perkins (1991), who study the same problem using lattice programming. Majumdar, Mitra and Nyarko (1989) study optimal growth for stochastic nonconvex models, as do Nishimura and Stachurski (2004). Dimaria and Le Van (2002) analyze the dynamics of deterministic models with R&D and corruption. [60]

6 Credit Markets, Insurance and Risk

In terms of informational requirements necessary for efficient free market operation and low transaction costs, one of the most problematic of all markets is the intertemporal trade in funds. Here information is usual asymmetric, and lenders face the risk of both voluntary and involuntary default (Kehoe and Levine 1993). Voluntary default is strategic default by borrowers who judge

[60] One should be cautious about interpreting these nonconvex models as aggregative studies of development. The Second Fundamental Welfare Theorem does not apply, so decentralization is problematic.

the expected rewards of repayment to be lower than those of not repaying the loan. Involuntary default occurs when ex post returns are insufficient to cover total loans.

Facing these risks, a standard response of lenders is to make use of collateral (Kiyotaki and Moore 1997). But the poor lack collateral almost by definition; as a result they are credit constrained. Credit constraints in turn restrict participation by the poor in activities with substantial set up costs, as well as those needing large amounts of working capital. For the poor, then, the range of feasible income-generating activities is reduced. Thus, the vicious circle of poverty: Income determines wealth and low wealth restricts collateral. This trap is discussed in Section 6.1. [61]

The market for insurance is similar to the market for credit, in that information is asymmetric and transaction costs are high. This can lead to poverty traps in several ways. In Section 6.2, we study a model where poor agents, lacking access to insurance or credit, choose low risk strategies at the cost of low mean income. These choices reinforce their poverty.

In Section 6.3 we review Matsuyama's (2004) world economy model, where all countries must compete for funds in a global financial market. On one hand, diminishing returns imply that rewards to investment in the poor countries are large. High returns attract funds and investment, and high investment provides a force for convergence. On the other hand, credit markets are imperfect, and rich countries have more collateral. This puts them in a strong position vis-a-vis the poor when competing for capital. The inability of the poor to guarantee returns with collateral is a force for divergence.

6.1 Credit markets and human capital

Consider an economy producing only one good and facing a risk free world interest rate of zero. Agents live for one period. Each has one and only one child. From their parent, the child receives a bequest x. At the beginning of life, each agent chooses between two occupations. The first is to work using a constant returns technology $Y = \bar{w}L$, where Y is output, L is total labor input in this sector, and w is a productivity parameter. The agent supplies all

[61] See also Tsiddon (1992) for a poverty trap model connected to the market for credit. In his model, asymmetric information leads to a moral hazard problem, which restricts the ability of investors to raise money. The market solution involves quantity constraints on loans, the severity of which depends on the level of income.

of his or her labor endowment ℓ_t, and we define $w_t := \bar{w}\ell_t$ as the return to this choice of occupation. We admit the possibility that ℓ_t varies stochastically, so w_t may be random.

Alternatively, the agent may set up a project at cost F. The gross payoff from the project is equal to Q_t. Agents with wealth $x_t < F$ may borrow to cover the costs of the project beyond which they are able to self-finance. They face interest rate $i > 0$, where the excess of the borrowing rate over the risk free rate reflects a credit market imperfection. In this case we have in mind costs imposed on lenders due to the need for supervision and contract enforcement (c.f., e.g., Galor and Zeira 1993, p. 39). These costs are then passed on to the borrower.

The two stochastic productivity parameters w_t and Q_t are draws from joint distribution φ. We assume that $\mathbb{E}\ell_t = 1$, and that $\mathbb{E}w_t = \bar{w} < \mathbb{E}Q_t - F$. Thus, the net return to setting up the project is higher on average than the wage. However, the agent may still choose to work at wage rate w_t if his or her income is relatively low. The reason is that for the poor setting up a project requires finance at the borrowing rate $i > 0$, which may offset the differential return between the two occupations.

Consider the employment decisions and wealth dynamics for each dynasty. Omitting time subscripts, an agent with bequest x has

$$y := \text{lifetime income } = \begin{cases} x + w & \text{if do not set up project;} \\ (x - F)(1 + i) + Q & \text{if set up project, } x < F; \\ (x - F) + Q & \text{if set up project, } x \geq F. \end{cases}$$

Preferences are given by $u(c, b) = (1 - \theta) \ln c + \theta \ln b$, where $\theta \in (0, 1)$ is a parameter, c is consumption and b is bequest to the child. As a result, each agent bequeaths a fraction θ of y; the remainder is consumed. Indirect utility is $v(y) = \gamma + \delta \ln y$, where $\gamma, \delta > 0$ are constants.

To abstract temporarily from the issue of risk aversion let us suppose that each agent can observe his or her idiosyncratic shocks (w_t, Q_t) prior to choosing a field of employment. As a result, agents with $x \geq F$ will choose to set up projects iff $Q - F \geq w$. Agents with $x < F$ will choose the same iff $(x - F)(1 + i) + Q \geq x + w$; in other words, iff

$$x \geq \hat{x} := \frac{w - Q + F(1 + i)}{i}.$$

It follows that dynamics for each dynasty's wealth in this economy are given

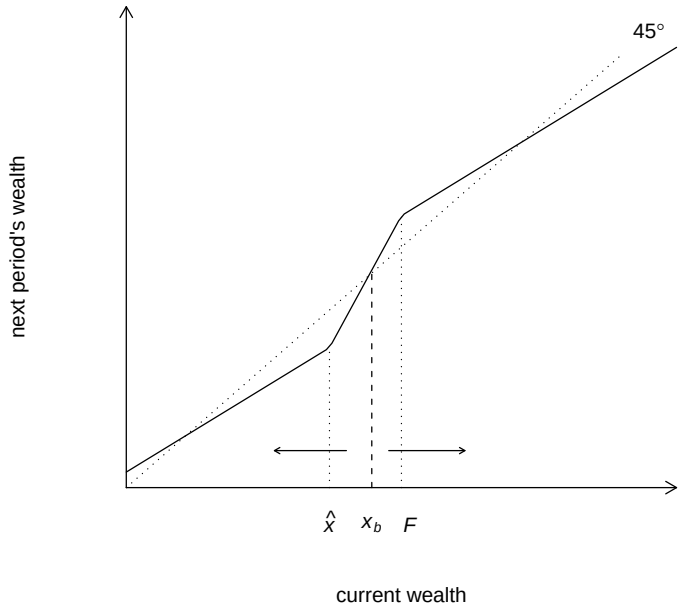


Fig. 19. Deterministic dynamics

by the transition rule

$$x_{t+1} = S_t(x_t); \quad S_t(x_t) = \theta \times \begin{cases} x_t + w_t & \text{if } x_t \leq \hat{x}_t; \\ (x_t - F)(1+i) + Q_t & \text{if } x_t \in (\hat{x}_t, F); \\ x_t - F + Q_t & \text{if } x_t \geq F. \end{cases}$$

Figure 19 illustrates a transition rule S and hence the dynamics of this economy when the two rates of return are constant and equal to their means.[62] For this particular parameterization there are multiple equilibria. Agents with initial wealth less than the critical value x_b will converge to the lower attractor, while those with greater wealth will converge to the high attractor. Given any initial distribution ψ_0 of wealth in the economy, the fraction of agents converging to the lower attractor will be $\int_0^{x_b} \psi_0$. If this fraction is large, average long run income in the economy will be small.

A more realistic picture can be obtained if the productivity parameters are permitted to vary stochastically around their means. This will allow at least some degree of income mobility—perhaps very small—which we tend to observe over time in almost all societies. To this end, suppose that for each agent

[62] The parameters here are set to $\theta = 0.7$, $w = 0.06$, $Q_t \equiv 1.05$, $i = 2$ and $F = 0.65$.

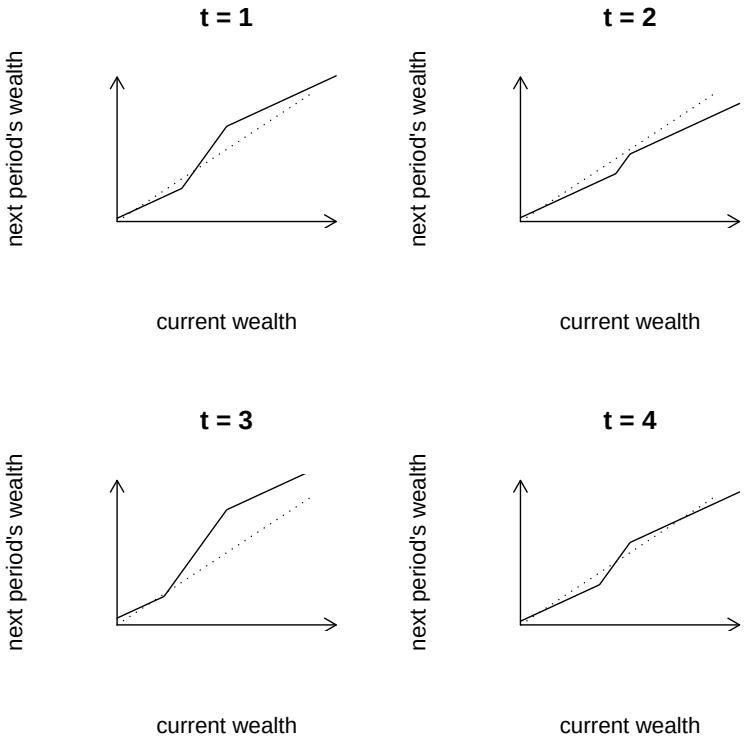


Fig. 20. Stochastic law of motion

and at each point in time the parameters w_t and Q_t are drawn independently across time and agents from a bivariate lognormal distribution. In this case the transition law is itself random, and varies for each agent at each point in time.

Figure 20 shows a simulated sequence of transition rules facing a given agent starting at $t = 1$. At $t = 2$ a negative shock to the project return Q causes the high level attractor to disappear. A *series* of such negative shocks would cause a rich dynasty to loose its wealth. In this case, however, the shocks are iid and such an outcome is unlikely. It turns out that the time 3 shocks are strongly positive.

If the number of agents is large, then the sequence of cross-sectional distributions for wealth over time can be identified with the sequence of marginal probability laws $(\psi_t)_{t\geq 0}$ generated by the Markov process $x_{t+1} = S_t(x_t)$. It is not difficult to prove that this Markov process is ergodic. The intuition and the dynamics are more or less the same as for the nonconvex growth model of Section 3.3. [63] We postpone further details on dynamics until the next section,

[63] As we discussed at length in that section, it would be a mistake to claim that this

which treats another version of the same model.

There are several interpretations of the two sector story with fixed costs described above. One is to take the notion of a project or business literally, in which case F is the cost of set up and working capital which must be paid up before the return is received. Alternatively F might be the cost of schooling, and Q is the payoff to working for skilled individuals.[64] As emphasized by Loury (1981) and others, human capital is particularly problematic for collateral-backed financing, because assets produced by investment in human capital cannot easily be bonded over to cover the risk of default.

Whatever the precise interpretation, the "project" represents an opportunity for the poor to lift themselves out of poverty, while the fixed cost F and the credit market imperfection captured here by i constitute a barrier to taking it. Microeconometric studies suggests that the effects of this phenomenon are substantial. For example, Barrett, Bezuneh and Aboud (2001) analyze the effects of a large devaluation of the local currency that occurred in Côte d'Ivoire in 1994 on rural households. They find that "A macro policy shock like an exchange rate devaluation seems to create real income opportunities in the rural sector. But the chronically poor are structurally impeded from seizing these opportunities due to poor endowments and liquidity constraints that restrict their capacity to overcome the bad starting hand they have been dealt." (Barrett et al. 2001, p. 12)

The same authors also study a local policy shock associated with food aid distribution in Keyna. According to this study, "The wealthy are able to access higher-return niches in the non-farm sector, increasing their wealth and reinforcing their superior access to strategies offering better returns. Those with weaker endowments ex ante are, by contrast, unable to surmount liquidity barriers to entry into or expansion of skilled non-farm activities and so remain

---

ergodicity result in some way overturns the poverty trap found in the deterministic version.

[64] For these and related stories see Ray (1990), Ray and Streufert (1993), Banerjee and Newman (1993), Galor and Zeira (1993), Ljungqvist (1993), Freeman (1996), Quah (1996), Aghion and Bolton (1997), Piketty (1997), Matsuyama (2000) Mookherjee and Ray (2003) and Banerjee (2003). Yet another possible interpretation of the model is that F is the cost of moving from a rural to an urban area in order to find work. In the presence of imperfect capital markets, such costs—interpreted broadly to include any extra payments incurred when switching to the urban sector—may help to explain the large and growing differentials between urban and rural incomes in some modernizing countries.

trapped in lower return...livelihood strategies." (Barrett et al. 2001, p. 15).

6.2 Risk

For the poor another possible source of historical self-reinforcement is risk. In the absence of well-functioning insurance and credit markets, the poor find ways to mitigate adverse shocks and to smooth out their consumption. One way to limit exposure is to pass up opportunities which might seem on balance profitable but are thought to be too risky. Another strategy is to diversify activities; and yet another is to keep relatively large amounts of assets in easily disposable form, rather than investing in ventures where mean return is high. All of these responses of the poor to risk have in common the fact that they tend to lower mean income and reinforce long run poverty.

A simple variation of the model from the previous section illustrates these ideas.[65] Let the framework of the problem be the same, but current shocks are no longer assumed to be previsible. In other words, each agent must decide his or her career path before observing the shocks w_t and Q_t which determine individual returns in each sector. Given that preferences are risk averse (indirect utility is $v(y) = \gamma + \delta \ln y$), the agent makes these decisions as a function not only of mean return but of the whole joint distribution. Regarding this distribution, we assume that both shocks are lognormal and may be correlated.

Lenders also cannot observe these variables at the start of time t, and hence the borrowing rate $i = i(x)$ reflects the risk of default, which in turn depends on the wealth x of the agent. In particular, default occurs when Q_t is less than the debtor's total obligations $(F - x)(1 + i(x))$. In that case the debtor pays back what he or she is able. Lifetime income is therefore

$$y = \begin{cases} x + w & \text{if do not set up project;} \\ \max\{0, (x - F)(1 + i(x)) + Q\} & \text{if set up project, } x < F; \\ (x - F) + Q & \text{if set up project, } x \geq F. \end{cases}$$

It turns out that in our very simplistic environment agents will never borrow, because when shocks are lognormal agents with $x < F$ who borrow will have $\mathbb{P}\{y = 0\} > 0$, in which case $\mathbb{E}v(y) = -\infty$. (If $x \geq F$ agents may still choose to work for a wage, depending on the precise joint distribution.) The result that agents never borrow is clearly unrealistic. For more sophisticated

[65] What follows is loosely based on Banerjee (2003).

versions of this model with similar dynamics see Banerjee (2003) or Checchi and García-Peñalosa's (2004).

Because agents never borrow, the dynamics for the economy are just

$$x_{t+1} = \theta(x_t + w_t) \cdot \mathbb{1}\{x_t \in D\} + \theta(x_t - F + Q_t) \cdot \mathbb{1}\{x_t \notin D\},$$

where $D := \{x : \mathbb{E}\, v(x + w_t) \geq \mathbb{E}\, v(x - F + Q_t)\}$. (As before, $\mathbb{1}$ is the indicator function.) The stochastic kernel Γ for this process can be calculated separately for the two cases $x \in D$ and $x \notin D$ using the same change-of-variable technique employed in Section 3.1. The calculation gives

$$\Gamma(x, x') = \varphi_w \left(\frac{x' - \theta x}{\theta} \right) \frac{1}{\theta} \cdot \mathbb{1}\{x \in D\} + \varphi_Q \left(\frac{x' - \theta(x - F)}{\theta} \right) \frac{1}{\theta} \cdot \mathbb{1}\{x \notin D\},$$

where φ_w and φ_Q are the marginal densities of w and Q respectively.

A two-dimensional plot of the kernel is given in Figure 21, where the parameters are $F = 1$, $\theta = 0.45$, $\ln w \sim N(0.1, 1)$, and $\ln Q \sim N(1.4, 0.2)$. The dark unbroken line is the 45^o line. Lighter areas indicate greater elevation, in this case associated with a collection of probability mass. For the parameters chosen, agents work precisely when $x < F$, and set up projects when $x \geq F$ (so that $D = [0, F]$), despite the fact that mean returns to the project are higher than those of working. The concentration of probability mass along the 45^o line in the region $D = [0, F]$ implies that poverty will be strongly self-reinforcing.

Nevertheless, lognormal shocks give poor individuals a non-zero probability of becoming rich at every transition; and the rich can eventually become poor, although it might take a sequence of negative shocks. The rate of mixing depends on the parameters that make up the law of motion and the variance of the shock. Usually some small degree of mixing is a more natural assumption than none. The mixing causes the corresponding Markov chain to be ergodic. This is the case regardless of how small the tails of the shocks are made. [66] For more details on ergodicity see the technical appendix.

To summarize, the poor are not wealthy enough to self-insure, and as a result choose income streams that minimize risk at the expense of mean earnings. The effect is to reinforce poverty. A number of country studies provide evidence of this behavior. [67] Dercon (2003) finds that the effects on mean income are substantial. In a review of the theoretical and empirical literature, he estimates

[66] But not necessarily so if the shocks have bounded support.
[67] See, for example, Morduch (1990) and Dercon (1998).

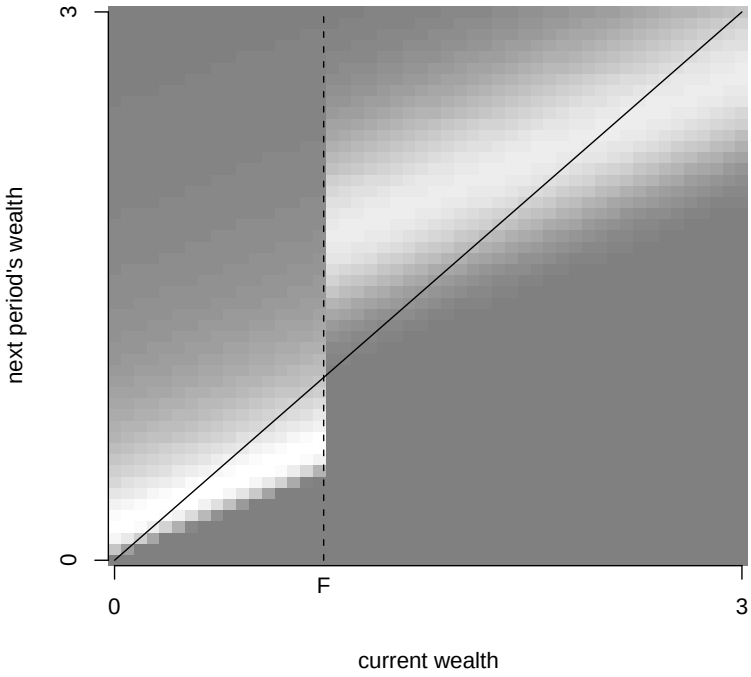


Fig. 21. Stochastic kernel

that incomes of the poor could be 25–50% higher on average if they had the same protection against shocks that the rich had as a result of their wealth (Dercon 2003, p. 14).

A more sophisticated model of the relationship between risk and development is Acemoglu and Zilibotti (1997). In their study, indivisibilities in technology imply that diversification possibilities are tied to income. An increase in investment raises output, which then improves the extent of diversification. Since agents are risk averse, greater diversification encourages more investment. In the decentralized outcome investment is too small, because agents do not take into account the effect of their investment on the diversification opportunities of others.

6.3 Credit constraints and endogenous inequality

Next we consider a world economy model with credit market imperfections due to Matsuyama (2004). For an individual country, the formulation of the problem is as follows. A unit mass of agents live for two periods each, supplying

one unit of labor in the first period of life and consuming all their wealth in the second. Per capita output of the consumption good is given by $y_t = f(k_t)\xi_t$, where f is a standard concave production function, k_t is the capital stock and $(\xi_t)_{t\geq 0}$ is a noise process. Once the current shock ξ_t is realized production then takes place. Factor markets are competitive, so that labor and capital receive payments $w_t = [f(k_t) - k_t f'(k_t)]\xi_t =: w(k_t, \xi_t)$ and $\varrho_t = f'(k_t)\xi_t$ respectively.

Current wages w_t are invested by young agents to finance consumption when old. Funds can be invested in a competitive capital market at gross interest rate R_{t+1}, or in a project which transforms one unit of the final good into Q units of the capital good at the start of next period. It is assumed that projects are discrete and indivisible: Each agent can run one and only one project. [68] They will need to borrow $1 - w_t$, the excess cost of the project over wages.

Our agents are risk neutral. Time t information is summarized by the information set $\mathscr{F}_t$, and we normalize $\mathbb{E}[\xi_{t+1} \,|\, \mathscr{F}_t] = 1$. In the absence of borrowing constraints, agents choose to start a project if $\mathbb{E}[\varrho_{t+1}Q - R_{t+1}(1 - w_t) \,|\, \mathscr{F}_t] \geq \mathbb{E}[R_{t+1}w_t \,|\, \mathscr{F}_t]$. This is equivalent to

$$\mathbb{E}[R_{t+1} \,|\, \mathscr{F}_t] \leq \mathbb{E}[\varrho_{t+1}Q \,|\, \mathscr{F}_t]. \tag{26}$$

However, it is assumed that borrowers can credibly commit to repay only a fraction λ of revenue $\varrho_{t+1}Q$. Thus $\lambda \in [0, 1]$ parameterizes the degree of credit market imperfection faced by borrowers in this economy. As a result, agents can start a project only when $\mathbb{E}[\lambda\varrho_{t+1}Q \,|\, \mathscr{F}_t]$ exceeds $\mathbb{E}[R_{t+1}(1 - w_t) \,|\, \mathscr{F}_t]$, the cost of funds beyond those which the agent can self-finance. In other words, when $w_t = w(k_t, \xi_t) < 1$, we must have

$$\mathbb{E}[R_{t+1} \,|\, \mathscr{F}_t] \leq \Lambda(k_t, \xi_t)\mathbb{E}[\varrho_{t+1}Q \,|\, \mathscr{F}_t], \tag{27}$$

where $\Lambda(k_t, \xi_t) := \lambda/(1 - w_t)$ Given the profitability constraint (26), the borrowing constraint (27) binds only when $\Lambda(k_t, \xi_t) < 1$. [69]

In the case of autarky it turns out that adjustment of the domestic interest rate can always equilibrate domestic savings and domestic investment. Since each generation of agents has unit mass, total domestic savings is just w_t. If $w_t \geq 1$, then all agents run projects and total output of the capital good is

[68] Put differently, we imagine that output is Q units of capital good for all investment levels greater than or equal to one. See the original model for a more general technology.

[69] Of course if $w_t \geq 1$ then all agents can self-finance and the borrowing constraint never binds.

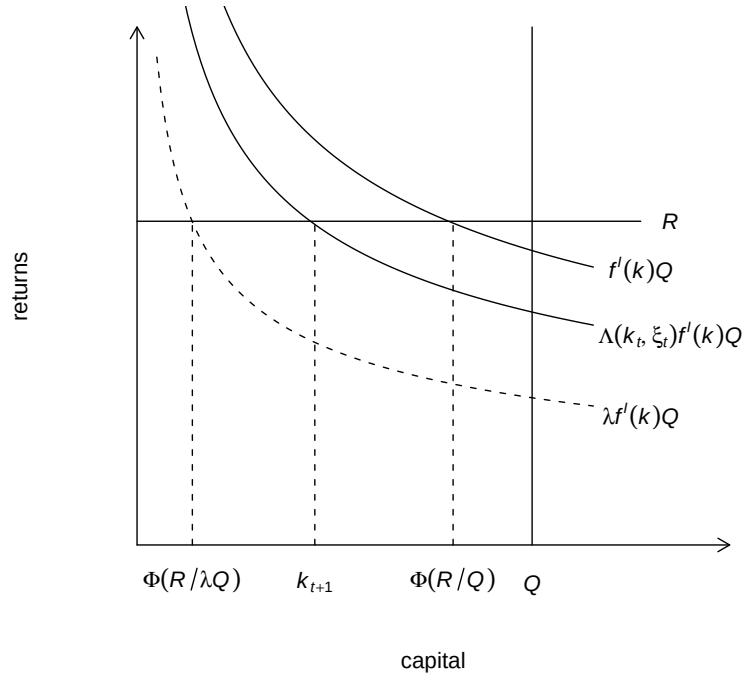


Fig. 22. Domestic investment

Q. If $w_t < 1$, then w_t is equal to the fraction of agents who can start projects. Output of the capital good is $w_t Q$. Assuming that capital depreciates totally in each period, we get $k_{t+1} = \min\{w(k_t, \xi_t)Q, Q\}$. If, for example, technology in the final good sector is Cobb-Douglas, so that $f(k) = Ak^\alpha$, where $\alpha < 1$, then $w(k_t, \xi_t) = (1-\alpha)Ak^\alpha \xi_t$. For $\xi_t \equiv 1$ there is a unique and globally stable steady state k^*.

A more interesting case for us is the small open economy. Here a world interest rate of R is treated as fixed and given. The final good is tradable, so international borrowing and lending are allowed. However, the project must be run in the home country (no foreign direct investment) and factors of production are nontradable.

In the open economy setting there is a perfectly elastic supply of funds at the world interest rate R. The effective demand for funds on the part of domestic projects is determined by (26) and (27). The right hand side of (26) is the expected marginal product of capital in this sector, $\mathbb{E}[\varrho_{t+1}Q \,|\, \mathscr{F}_t]$. Since $\mathbb{E}[\xi_{t+1}|\mathscr{F}_t] = 1$ we have $\mathbb{E}[\varrho_{t+1}Q \,|\, \mathscr{F}_t] = f'(k_{t+1})Q$. Absent borrowing constraints, investment adjusts to equalize $f'(k_{t+1})Q$ with R. Figure 22 shows the intersection of the curve $k \mapsto f'(k)Q$ with the horizontal supply curve R at

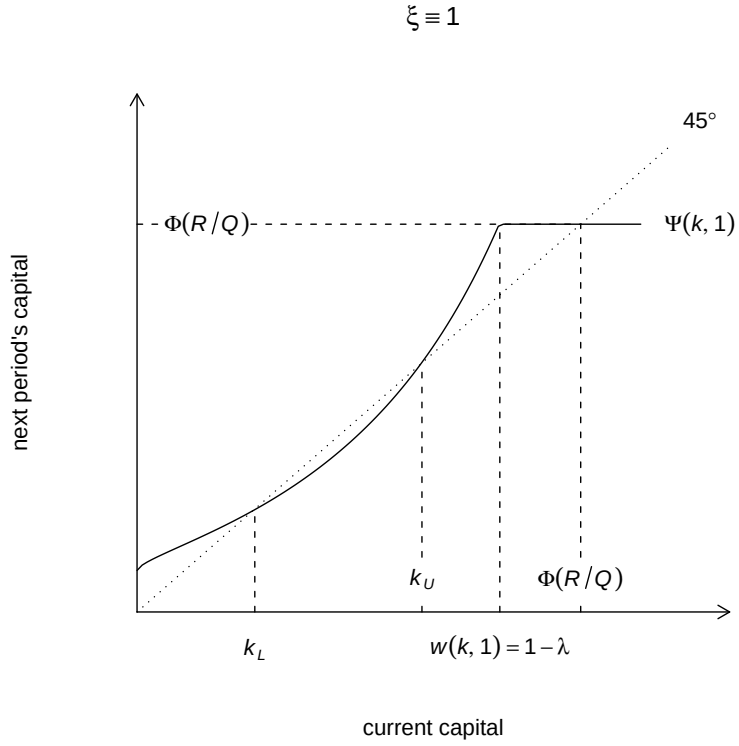


Fig. 23. Deterministic dynamics

$\Phi(R/Q)$, where Φ is the inverse function of f'.

As the figure is drawn, however, $\Lambda(k_t, \xi_t) < 1$, perhaps because the capital stock is small, or because of an adverse productivity shock. As a result, the borrowing constraint is binding, and next period's capital stock k_{t+1} is given by the intersection of the effective demand curve $k \mapsto \Lambda(k_t, \xi_t) f'(k) Q$ and the supply curve R.

Assuming that $\Phi(R/Q) < Q$ as drawn in the figure, the law of motion for the capital stock is $k_{t+1} = \Psi(k_t, \xi_t)$, where

$$\Psi(k, \xi) := \begin{cases} \Phi[R/\Lambda(k, \xi)Q] & \text{if } w(k, \xi) < 1 - \lambda; \\ \Phi(R/Q) & \text{if } w(k, \xi) \geq 1 - \lambda. \end{cases} \tag{28}$$

For $w(k_t, \xi_t) < 1 - \lambda$ we have $\Lambda(k_t, \xi_t) < 1$ and the borrowing constraint binds. Domestic investment is insufficient to attain the unconstrained equilibrium $\Phi(R/Q)$. In this region of the state space, the law of motion $k \mapsto \Psi(k, \xi)$ is increasing in k. Behind this increase lies a credit multiplier effect: Greater domestic investment increases collateral, which alleviates the borrowing constraint. This in turn permits more domestic investment, which increases col-

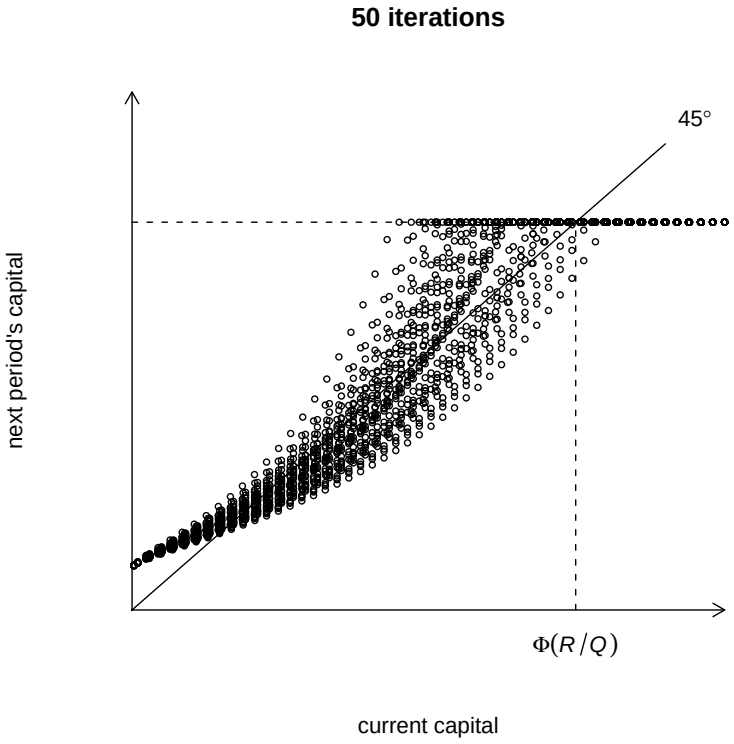


Fig. 24. Stochastic dynamics

lateral, and so on. Individual agents do not take into account the effect of their actions on the borrowing constraint.

Figure 23 shows the law of motion when $\xi_t \equiv 1$. As drawn, there is a poverty trap at k_L and another attractor at $\Phi(R/Q)$. Countries with $k_t > k_U$ tend to $\Phi(R/Q)$, while those with $k_t < k_U$ tend to k_L. Figure 24 shows stochastic dynamics by superimposing the first 50 laws of motion from a simulation on the 45^o diagram. The shocks $(\xi_t)_{t\geq 0}$ are independent and identically distributed. [70] Notice that for particularly good shocks the lower attractor k_L disappears, while for particularly bad shocks the higher attractor at $\Phi(R/Q)$ vanishes.

Figure 25 shows a simulated time series for the same parameters as Figure 24 over 400 periods. At around $t = 290$ the economy transitions to the higher attractor $\Phi(R/Q)$. Subsequent fluctuations away from this equilibrium are due to shocks so negative that $\Phi(R/Q)$ ceases to be an attractor (see Figure 24).

The story does not end here. What is particularly interesting about Matsuyama's study is his analysis of symmetry-breaking. He shows the following

[70] The production function is $f(k) = k^\alpha$. The shock is lognormal. The parameters are $\alpha = 0.59$, $Q = 2.4$, $\lambda = 0.40$, $R = 1$ and $\ln \xi \sim N(0.01, 0.08)$.

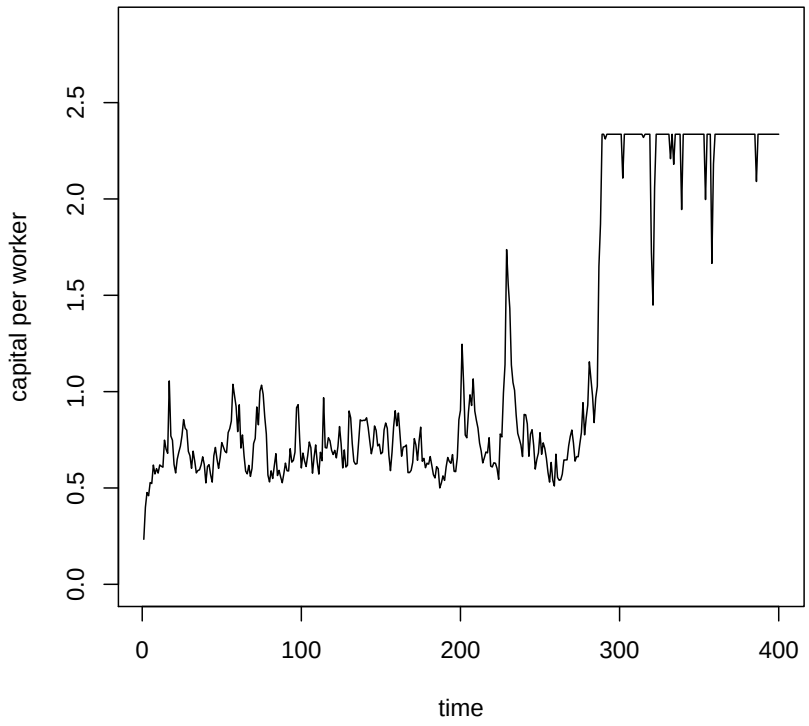


Fig. 25. Time series

for a large range of parameter values: For a world economy consisting of a continuum of such countries, the deterministic steady state for autarky, which is k^* defined by $k^* = w(k^*, 1)Q$, is precisely k_U, the *unstable* steady state for each country under open international financial markets and a world interest rate that has adjusted to equate world savings and investment. Figure 26 illustrates the situation.

Thus, the symmetric steady state after liberalization, where each country has capital stock k^*, is unstable and cannot be maintained under any perturbation. The reason is that countries which suffer from bad (resp., good) shocks are weakened (resp., strengthened) in terms of their ability to guarantee returns on loans, and therefore to compete in the world financial market. This leads to a downward (resp. upward) spiral. Under these dynamics the world economy is polarized *endogenously* into rich and poor countries.

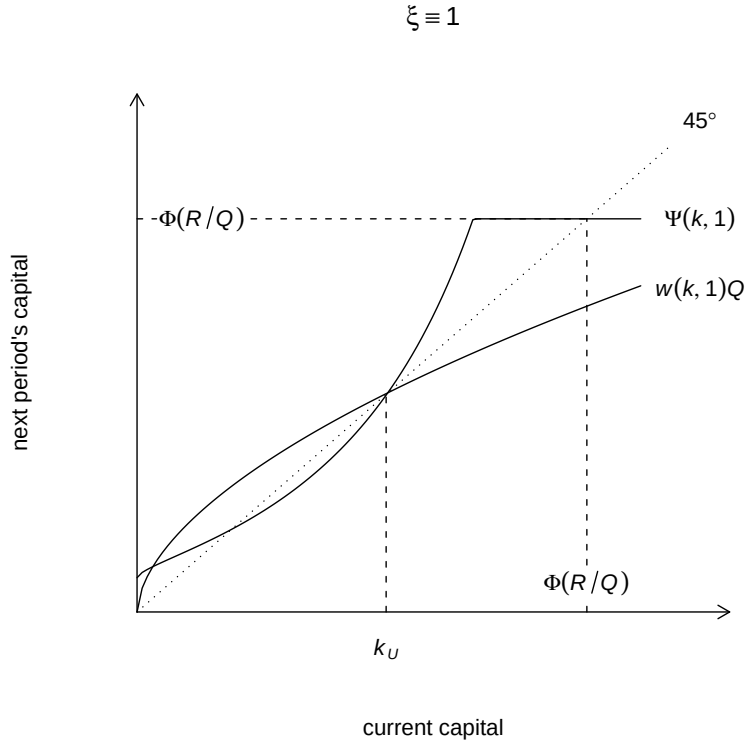


Fig. 26. Symmetry-breaking

7 Institutions and Organizations

The fundamental economic problem is scarcity. Since the beginning of life on earth, all organisms have engaged in competition for limited resources. The welfare outcomes of this competition have ranged from efficient allocation to war, genocide and extinction. It is the *rules of the game* which determine the social welfare consequences. More precisely, it is the long run interaction between the rules of the game and the agents who compete.

Institutions—which make up the rules of the game—were at one time thought to have strong efficiency properties in equilibrium. To a large extent, this is no longer the case (for an introduction to the literature, see, for example North 1993, 1995; or Hoff 2000). Institutions can either reinforce market failure or themselves be the source of inefficiency. Moreover, institutions are path dependent, so that bad equilibria forming from historical accident may be locked in, causing poverty to persist.

Among the set of institutions, the state is one of the most important determinants of economic performance; and one of the most common kinds of

"government failure" is corruption.[71] In Section 7.1 we review why corruption is thought to be not only bad for growth and development, but also self-reinforcing.

Section 7.2 then looks at the kinship system, a kind of institution that arises spontaneously in many traditional societies to address such market problems as lack of formal insurance. We consider how these systems may potentially form a local poverty trap, by creating hurdles to adoption of new techniques of production. Although the aggregate outcome is impoverishing, it is shown that the kinship system may nevertheless fail to be dismantled as a result of individual incentives.

7.1 Corruption and rent-seeking

Corruption is bad for growth. A number of ways that corruption retards development have been identified in the literature. First, corruption tends to reduce the incentive to invest by decreasing net returns and raising uncertainty. This effect impacts most heavily on increasing returns technologies with large fixed costs. Once costs are sunk, investors are subject to hold-up by corrupt officials, who can extort large sums. Also, governments and officials who have participated in such schemes find it difficult to commit credibly to new infrastructure projects.

Second, corruption diverts public expenditure intended for social overhead capital. At the same time, the allocation of such capital is distorted, because officials prefer infrastructure projects where large side payments are feasible. Corruption also hinders the collection of tax revenue, and hence the resource base of the government seeking to provide public infrastructure. Again, a lack of social overhead capital such as transport and communication networks tends to impact more heavily on the modern sector.

Third, innovators suffer particularly under a corrupt regime, because of their higher need for such official services as permits, patents and licenses (De Soto 1989; Murphy, Shleifer and Vishny 1993). The same is true for foreign investors, who bring in new technology. Lambsdorff (2003) finds that on average

[71] Following the excellent survey of Bardhan (1997), we define corruption to be "the use of public office for private gains, where an official (the agent) entrusted with carrying out a task by the public (the principal) engages in some sort of malfeasance for private enrichment which is difficult to monitor for the principal" (Bardhan 1997, p. 1321).

a 10% worsening in an index of transparency and corruption he constructs leads to a fall of 0.5 percentage points the ratio of foreign direct investment to GDP.

Not only is corruption damaging to growth, but it also tends to breed more corruption. In other words, there are complementarities in corruption and other rent-seeking activities. It is this increasing returns nature of corruption which may serve to lock in poverty. Some equilibria will be associated with high corruption and low income, where many rent-seekers prey on relatively few producers. Others will have the reverse.

The decision of one official to seek bribes will increase expected net rewards to bribe taking in several ways. The most obvious of these complementarities is that when many agents are corrupt, the probability of detection and punishment for the marginal official is lowered. A related point is that if corruption is rampant then detection will not entail the same loss of reputation or social stigma as would be the case in an environment where corruption is rare. In other words, corruption is linked to social norms, and is one of the many reasons why they matter for growth. [72] Third, greater corruption tends to reduce the search cost for new bribes.

Murphy, Shleifer and Vishny (1993) point out yet another source of potential complementarities in rent-seeking. Their idea is that even if returns to predation are decreasing in an absolute sense, they may still be increasing *relative* to production. This would occur if the returns to productive activities—the alternative when agents make labor supply decisions—fall *faster* than those to rent-seeking as the number of rent-seekers increases. The general equilibrium effect is that greater rent-seeking decreases the (opportunity) cost of an additional rent-seeker.

In their model there is a modern sector, where output by any individual is equal to a, and a subsistence technology with which agents can produce output $c < a$. Alternatively, agents can prey on workers, obtaining for themselves an amount no more than b per person, but limited by the amount of output available for predation. This in turn depends on the number of people working in the productive sectors. The authors assume, in addition, that only modern

[72] Transparency International's 2004 Global Corruption Report cites a statement by the president of the Government Action Observatory in Burundi that "corruption has spread, openly and publicly, to such an extent that those who practice it have become stronger than those who are fighting against it. This has led to a kind of *reversal of values*." (Emphasis added.)

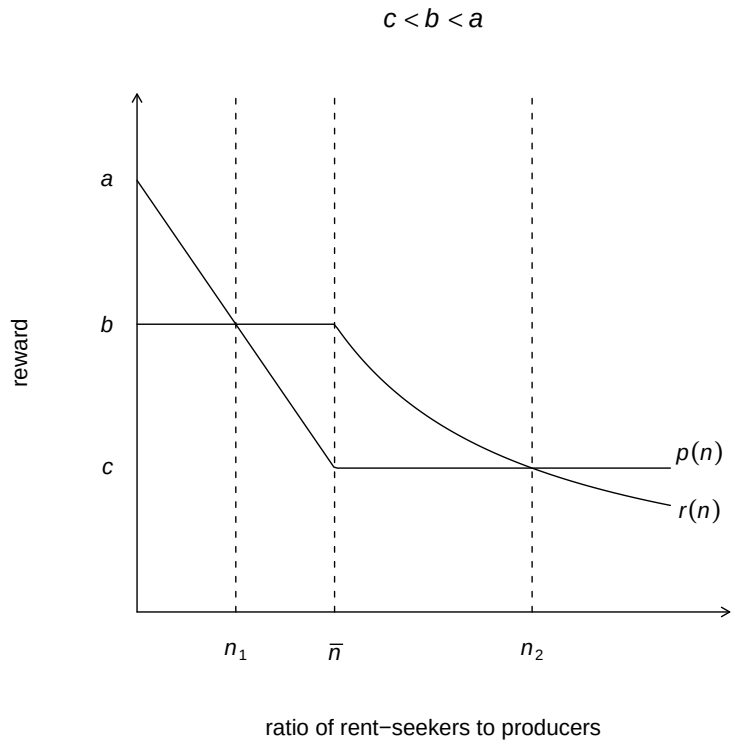


Fig. 27. Rewards to rent-seeking and production

sector output can be appropriated by rent-seekers, so returns to subsistence farming are always equal to c.

An equilibrium is an allocation of labor across the different occupations such that returns to all are equal, and no individual agent can increase their reward by acting unilaterally. To locate equilibria, we now discuss returns to working in the different sectors as a function of n, which is defined to be the number of rent seekers for each modern sector producer.

Returns to employment in the subsistence sector are always given by c. Rent-seekers all take a slice b of the pie until their ratio to modern sector producers n satisfies $a - bn = c$. At this ratio, which we denote $\bar{n}$, the earnings of the modern sector producers fall to that of the subsistence producers, and the rent-seekers must reduce the size of their take (or earn nothing). After $\bar{n}$, the rent-seekers each take $(a - c)/n$, exactly equalizing returns to modern sector production and subsistence.

Let $p(n)$ and $r(n)$ be returns to modern sector production and rent-seeking respectively, so that $p(n) = (a - bn)\mathbb{1}\{n < \bar{n}\} + c\mathbb{1}\{n \geq \bar{n}\}$ and $r(n) = b\mathbb{1}\{n < \bar{n}\} + \frac{a-c}{n}\mathbb{1}\{n \geq \bar{n}\}$. These curves are drawn in Figure 27. The figure shows

that there are multiple equilibria whenever the parameters satisfy $c < b < a$. One is where all work in the modern sector. Then $n = 0$, and $p(n) = p(0) = a > r(n) = b > c$. This allocation is an equilibrium, where all agents earn the relatively high revenue available from modern sector production. In addition, because $b > c$, the payoff functions $n \mapsto p(n)$ and $n \mapsto r(n)$ intersect above $\bar{n}$, at n_2. This is again an equilibrium, where the payoffs to working in the subsistence sector, the modern sector and the rent-seeking sector are all equal and given by c.

Notice that b does not affect income in either of these two equilibria. However, it does affect which one is likely to prevail. If b declines below c, for example, then only the good equilibrium will remain. If it increases above a, then the bad equilibrium will be unique. When there are two equilibria, higher b increases the basin of attraction for the bad equilibrium under myopic Marshallian dynamics.

In summary, the model exhibits a general equilibrium complementarity to corruption, which helps illustrate why corruption tends to be self-reinforcing, therefore causing poverty to persist. These kind of stories are important, because in practice corruption and related crimes tend to show a great deal of variation across time and space, often without obvious exogenous characteristics that would cause such variation.

There are many other models which exhibit self-reinforcement and path dependence in corruption. One is Tirole (1996), who studies the evolution of individual and group reputation. In his model, past behavior provides information about traits, such as honesty, ability and diligence. However, individual behavior is not perfectly observed. As a result, actions of the group or cohort to which the individual belongs have predictive power when trying to infer the traits of the individual. It follows that outcomes and hence incentives for the individual are affected by the actions of the group.

In this case we can imagine the following scenario. Young agents progressively joint an initial cohort of workers, a large number of whom are known to be corrupt. Because the behavior of new agents is imperfectly observed, they inherit the suspicion which already falls on the older workers. As a result, they may have little incentive to act honestly, and drift easily to corruption. This outcome in turn perpetuates the group's reputation for corrupt action.

One can contemplate many more such feedback mechanisms. For example, it is often said that the low wages of petty officials drive them to corruption. But if corruption lowers national output and hence income, then this will reduce

the tax base, which in turn decreases the amount of resources with which to pay wages. For further discussion of corruption and poverty traps see Bardhan (1997). [73]

7.2 Kinship systems

All countries and economies are made up of people who at one time were organized in small tribes with their own experiences, customs, taboos and conventions. Over time these tribes were united into cities, states and countries; and the economies within which they operated grew larger and more sophisticated. Some of these economies became vibrant and strong. Others have stagnated. According to North (1993, p. 4),

> The reason for differing success is straightforward. The complexity of the environment increased as human beings became increasingly interdependent, and more complex institutional structures were necessary to capture the potential gains from trade. Such evolution required that the society develop institutions that will permit anonymous, impersonal exchange across time and space. But to the extent that "local experience" had produced diverse mental models and institutions with respect to the gains from such cooperation, the likelihood of creating the necessary institutions to capture the gains from trade of more complex contracting varied.

North and other development thinkers have emphasized that success depends on institutions rewarding efficient, productive activity; and having sufficient flexibility to cope with the structural changes experienced in the transition to modernity. The degree of flexibility and ability to adapt determines to what extent an economy can take advantage of the application of science, of new techniques, and of specialization and the effective division of labor.

To illustrate these ideas, in this section we review recent analysis of the "kin" system, an institution found in many traditional societies, usually defined as an informal set of shared rights and obligations between extended family and friends for the purpose of mutual assistance. [74] Where markets and state institutions are less developed, the kin system replaces formal insurance and

[73] For other kinds of poverty traps arising through interactions between the state and markets, see, for example, Hoff and Stiglitz (2004), or Gradstein (2004).
[74] A related form of local poverty traps is those generated by neighborhood effects. See Durlauf (2004).

social security by implementing various forms of community risk sharing, and by the provision of other social services (Hoff and Sen 2004). The question we ask in the remainder of this section is how, in the process of development, the kin system interacts with the nascent modern sector, and whether or not it may serve to *impede* the diffusion of new technologies and the exploitation of gains from trade.

An interesting example of such analysis is Baker (2004), who interprets Africa's lack of robust growth as a failure of technology diffusion caused by institutional barriers. She presents a model of a rural African village, and suggests two path dependent mechanisms related to the kin system which may serve to retard growth. Both of them involve community risk sharing, and indicate how technology adoption may have positive network externalities beyond simple social learning.

The first mechanism concerns risk sharing among kin members in the form of interest free "loans" with no fixed repayment schedule. Kin members in need can expect to receive these transfers from the better off, who in turn must comply or face various social sanctions (including, in the countries Baker studied, accusations of witchcraft as the source of their good fortune). Beyond the obvious incentive effects on those who might seek to improve their circumstances by using new technology, Baker suggests that a kin member who adopts new techniques may face significant additional uncertainty vis-a-vis income net of transfers if the kin group makes mistakes in estimating his or her true profits. Such a miscalculation may lead to excessive demands for "gifts" or other transfers.

As Baker points out, the uncertainty effect of the transfers will be larger for those who adopt new technology, where costs and revenue are harder for the other kin to estimate. For example, the kin may have difficulty in measuring the real costs of new techniques, such as fertilizer or more expensive seed, causing them to overestimate true profits. (New techniques are often associated with higher revenues combined with higher costs.)

On the other hand, cost and net profit will be easier to estimate if more kin members have experience of the new techniques. In other words, uncertainty will be mitigated for the marginal adopter if more of his or her fellow kin members adopt the same technology. As a result there are positive network externalities in terms of expected cost. This mechanism generates a coordination problem, whereby a critical mass of co-adopters may be necessary to make the new technology more attractive than the old. This need for coordination

may present a barrier to adoption.

At the same time, the coordination barrier would not seem to be insurmountable. Perhaps a kin group can negotiate to a better equilibrium when the gains are genuinely large? Baker suggests that in fact this will not be easy, because the risk sharing problem interacts with other path dependent institutions.

One of these concerns the nature of old age insurance among self-employed African farmers. Given the lack of state pensions and the difficulty of accumulating assets, support in old age may be contingent on the old providing some form of useful service to the household from which resources are to be acquired. And the most likely candidate for productive service from elderly farmers is the benefit of their experience. The problem here is that the value of this service provided by the old depends on a stagnant technology which does not change from generation to generation. Under new techniques the experience of old farmers may become redundant. If old farmers are able to resist the introduction of new techniques then it will be in their interests to do so. Once again, this is a source of multiple equilibria. The reason is that if the newer technology were already adopted then presumably it would be supported by old farmers, because this is then the methodology in which they have experience.

Another interesting study of the kin system has been conducted recently by Hoff and Sen (2004). They analyze the migration of kin members from rural areas to modern sector jobs, and show how network externalities arise in the migration decision. Even if kin members can coordinate on simultaneous migration, Hoff and Sen suggest that the kin group may put up barriers to prevent the loss of their most productive members. It is shown that even when the kin decisions are made by a majority, the barriers can be inefficient in terms of aggregate group welfare.

A simplified version of their story runs as follows. Kin members who do migrate may find themselves besieged by their less fortunate brethren. The latter come seeking not only "gifts" of cash transfers, but also help in finding jobs in the modern sector for themselves. Realizing this, employers will find it profitable to restrict employment of kin members. Here we assume these barriers are so high that migration while maintaining kin ties is never optimal. As a result, kin members choose between remaining in the rural sector or migrating while breaking their kin ties.

The kin group is thought of as a continuum of members with total mass of one. A fraction $\bar{\alpha} \in (0, 1)$ of the kin receive job offers in the modern sector.

The utility of remaining in the rural sector is

$$u_s(\alpha) = s_0 + b(1 - \alpha), \tag{29}$$

where here and elsewhere $\alpha \leq \bar{\alpha}$ is the fraction of the kin who break ties and move. The constant s_0 is a stand-alone payoff to rural occupation. The constant b is positive, so that utility of staying is higher when more kin members remain. On the other hand, the utility of moving to the modern sector is

$$u_m(\alpha) = m_0 - c(1 - \alpha), \tag{30}$$

where m_0 is a payoff to working in the modern sector and c is a positive constant. The function $\alpha \mapsto c(1 - \alpha)$ is the cost of ending kin membership (measured in the utility equivalent of various social sanctions which we will not describe). It is assumed that the cost of breaking kin ties for the marginal kin member decreases as more members leave the kin group and shift to the modern sector. [75]

Consider the interesting case, where $u_m(0) < u_s(0)$ and $u_m(\bar{\alpha}) > u_s(\bar{\alpha})$. A pair of curves for (29) and (30) which fit this pattern are depicted in Figure 28. If no kin members take modern sector jobs then it is not optimal to do so for an individual member. On the other hand, if all those with offers take up jobs, then their utility payoff will be higher than the payoff of those who remain.

If, as in the figure, we also have $u_m(\bar{\alpha}) > u_s(0)$, then it seems plausible that the kin members with job offers will coordinate their way to the equilibrium where all simultaneously move to modern sector jobs. Kin groups are not as diffuse as some other groups of economic actors, and coordination should prove correspondingly less problematic.

However, Hoff and Sen show that when kin members are heterogenous, a majority may take steps to forestall coordination by the productive critical mass on movement to the modern sector. Moreover, they may do so even when this choice is inefficient in terms of the kin's aggregate group payoff. In doing so, the kin group becomes a "dysfunctional institution," responsible for enforcing an inefficient status quo.

[75] Hoff and Sen cite Platteau (2000), who writes that to leave and enter the modern sector, a kin member "needs the protection afforded by the deviant actions of a sufficient number of other innovators in his locality. Rising economic opportunities alone will usually not suffice to generate dynamic entrepreneurs in the absence of a *critical mass* of cultural energies harnessed towards countering social resistance..." (Emphasis added.)

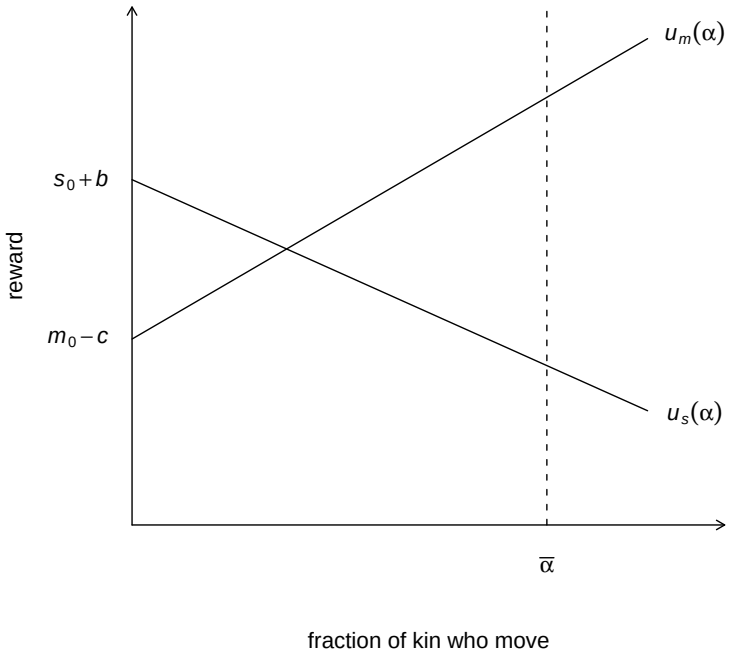


Fig. 28. Rewards to staying and moving

Their example works as follows. Consider a two stage game. First, the kin set the exit cost parameter c by majority vote. The two possible values are c^a and c^b, where $c^a < c^b$. Next, job offers are received, and kin members decide whether or not to move. Coordination always takes place in the situation where those with job offers together have a higher payoff in the modern sector.

There are now two types of kin members, those with high "ability" and those with low. The first type are of measure γ, and have probability α_H of getting a job offer from the modern sector. The second type are of measure $1 - \gamma$, and have probability α_L of getting a job offer from the modern sector, where $0 < \alpha_L < \alpha_H < 1$. We assume that $\gamma < 1/2$, so high ability types are in the minority. Also, we assume that $\gamma\alpha_H + (1-\gamma)\alpha_L = \bar{\alpha}$. Ex post, the law of large numbers implies that the fraction of kin members who get job offers will again be $\bar{\alpha}$.

Regarding parameters, we assume that $u_m^a(\alpha) := m_0 - c^a(1 - \alpha)$ satisfies $u_m^a(\bar{\alpha}) > u_s(0)$, but $u_m^b(\alpha) := m_0 - c^b(1-\alpha)$ satisfies $u_m^b(\bar{\alpha}) < u_s(0)$. The first inequality says that under the low cost regime, the payoff to working in the modern sector is greater than that of staying if all with job offers move. The second inequality says that under the high cost regime the opposite is true.

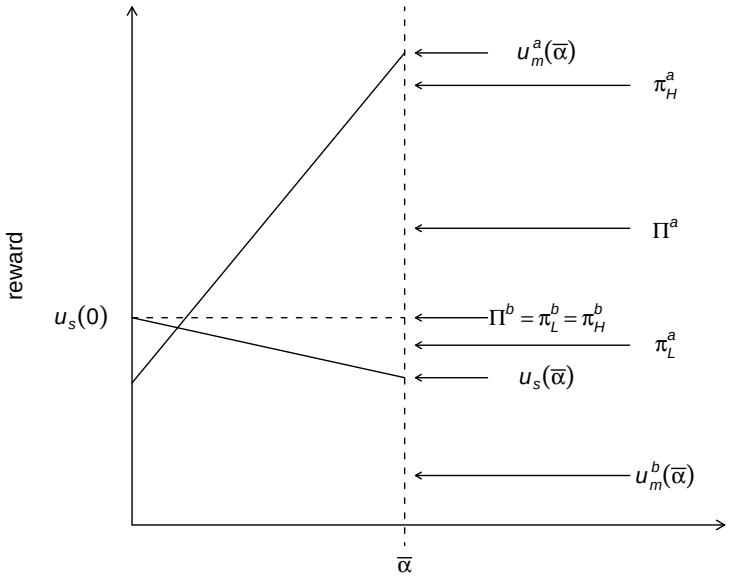


Fig. 29. Rewards

Because of coordination, under c^a all of those with job offers will move. The ex ante payoff of the high ability types is

$$\pi_H^a := \alpha_H u_m^a(\bar{\alpha}) + (1 - \alpha_H) u_s(\bar{\alpha}),$$

while that of low ability types is

$$\pi_L^a := \alpha_L u_m^a(\bar{\alpha}) + (1 - \alpha_L) u_s(\bar{\alpha}).$$

Under c^b all remain in the traditional sector, so the payoffs are $\pi_L^b := u_s(0) =: \pi_H^b$. Ex ante aggregate welfare measured as the sum of total payoffs is given under c^a by

$$\Pi^a := \bar{\alpha} u_m^a(\bar{\alpha}) + (1 - \bar{\alpha}) u_s(\bar{\alpha}).$$

Under c^b it is $\Pi^b := u_s(0)$.

What Hoff and Sen point out is that under some parameters it is possible to have

$$\pi_L^a < u_s(0) = \pi_L^b = \pi_H^b = \Pi^b < \Pi^a < \pi_H^a. \tag{31}$$

In this case $\pi_L^a < u_s(0) = \pi_L^b$, and since those with low ability are in the majority they will choose to set $c = c^b$. But then aggregate welfare is reduced,

because $\Pi^b < \Pi^a$. This situation is illustrated in Figure 29.[76] Incentives are such that the kinship institution perpetuates a low average income status quo.

8 Other Mechanisms

The poverty trap literature is vast, and even in a survey of this size many models must be neglected. A few of the more egregious omissions are listed in this section.

One of the earliest streams of literature on poverty traps is that related to endogenous fertility. A classic contribution is Nelson (1956), who shows how persistent underdevelopment can result from demographics. In his model, any increase in income lowers the death rate, which increases population and lowers capital stock per worker. If the population effect is stronger than diminishing returns then capital per worker cannot rise. See Azariadis (1996, Section 3.4) for other mechanisms and more references.

Other kinds of traps that arise in convex economies with complete markets include impatience traps and technology traps. Impatience traps typically involve subsistence levels of consumption, and sensitivity of consumption to income at low levels. See Magill and Nishimura (1984) or Azariadis (1996, Section 3.1). Technology traps are associated with low degrees of substitutability between capital and labor. See Azariadis (1996, Section 3.2).

See Dasgupta and Ray (1986) or Dasgupta (2003) for an introduction to the literature on malnutrition and underdevelopment. See also Basu and Van (1998) for a model of child labor with multiple equilibria.

9 Conclusions

The poor countries are not rich because they have failed to adopt the modern techniques of production which first emerged in Britain during the Industrial Revolution and then spread to some other nations in Western Europe and elsewhere. As a result, their economies have stagnated. By contrast, the rich countries possess market environments where the same techniques have been

[76] The parameters are $s_0 = 0.8$, $b = 0.2$, $m_0 = 2$, $c^a = 1.1$, $c^b = 2.3$, $\alpha_H = 0.9$, $\alpha_L = 0.1$ and $\gamma = 0.45$.

continuously refined, upgraded and extended, leading to what are now striking disparities between themselves and the poor.

Why would techniques not be adopted even when they are more efficient? Is it not the case that more efficient techniques are more profitable? The main objective of this survey has been to review a large number of studies which show why self-reinforcing traps may prevent the adoption of new technologies. For example, Section 5 showed how increasing returns can generate an incentive structure whereby agents avoid starting modern sector businesses, or invest little in their own training. Section 6 focused on credit market imperfections. Poor individuals lack collateral, which restricts their ability to raise funds. As a result, projects with large fixed costs are beyond the means of the poor, leaving them locked in low return occupations such as subsistence farming.

Recently many economists have highlighted the role of institutions in perpetuating poverty. Section 7 looked at why rent-seeking is both bad for growth and yet strongly self-reinforcing. Essentially similar societies may exhibit very different levels of predation simply as a result of historical accident, or some spontaneous coordination of beliefs. In addition, the role of kinship systems was analyzed as representative of the kinds of social conventions which may potentially harm formation of the modern sector.

Together, these mechanisms add up to a very different picture of development than the convex neoclassical benchmark model on which so much of modern growth theory has been based. Growth is not automatic. Small initial differences are magnified and then propagated through time. Poverty coexists with riches, much as it is observed to do in the cross-country income panel.

9.1 Lessons for economic policy

There is a real sense in which poverty trap models are optimistic. Poverty is not the result of some simple geographic or cultural determinism. The poor are not condemned to poverty by a set of unfavorable exogenous factors, or even a lack of resources. Temporary policy shocks will have large and permanent effects if one-off interventions can cause the formation of new and better equilibria.

In practice, however, engineering the emergence of more efficient equilibria seems problematic for a number of reasons. First, we have seen many examples of how bad equilibria can be stable and self-reinforcing. In this case small policy changes are not enough to escape from their grip. Large changes must

be made to the environment that people face, and the structure of their incentives. Such changes may be resisted by the forces that have perpetuated the inefficient equilibrium, such as a corrupt state apparatus fighting to preserve the status quo.

Second, coordinating changes in expectations and the status quo is difficult because norms and conventions are highly persistent. While it is possible to change policy and legislation almost instantaneously, it needs to be remembered that informal norms and conventions are often more important in governing behavior than the formal legalistic ones. Informal norms cannot be changed in the manner of interest rates, say, or tariffs. Rather they are determined within the system, and perpetuated by those forces that made them a stable part of the economy's institutional framework.

Third, policies can create new problems as a result of perverse incentives.[77] Successful policies will need to be carefully targeted, and operate more on the level of incentives than compulsion. These kinds of policies require a great deal of information. Traps which prevent growth and prosperity cannot be overcome without proper understanding and the careful design of policy.

A Technical Appendix

Section A.1 gives a general discussion of Markov chains and ergodicity. The proof of Proposition 3.1 is outlined. Section A.2 gives remaining proofs.

A.1 Markov chains and ergodicity

In the survey we repeatedly made use of a simple framework for treating Markov chains and ergodicity. The following is an elementary review. Our end

[77] For example, in South Korea the state is generally credited with solving many of the coordination problems associated with industrialization in that country through their organization and support of large industrial conglomerates, and through active policy-based lending. However, these actions also led to a moral hazard problem, as the industrial groups became highly leveraged with government-backed loans. In the 1970s, investment was increasingly characterized by a costly combination of duplication and poor choices. Losses were massive, and motivated subsequent liberalization.

objective is to sketch the proof of Proposition 3.1, but the review is intended to be more generally applicable.

Consider first a discrete time dynamical system evolving in state space $S \subset \mathbb{R}^n$. Just as for deterministic systems on S, which are represented by a *transition rule* associating each point in S with another point in S—the value of the state next period—a Markov chain is represented by a rule associating each point in S with a probability distribution over S. From this conditional distribution (i.e., distribution conditional on the current state $x \in S$) the next period state is drawn. In what follows the conditional distribution will be denoted by $\Gamma(x, dy)$, where $x \in S$ is the current state.

Because for Markov chains points in S are mapped into probability distributions rather than into individual points, it seems that the analytical methods used to study the evolution of these processes must be fundamentally different to those used to study deterministic discrete time systems. But this is not the case: Markov chains can always be reduced to deterministic systems.

To see this, note that since the state variable x_t is now a random variable, it must have some (marginal) distribution on S, which we call ψ_t. Suppose, as is often the case in economics, that ψ_t is a density on S, and that the distribution $\Gamma(x, dy)$ is in fact a density $\Gamma(x, y)dy$ for every $x \in S$. In that case the marginal distribution for x_{t+1} is a density ψ_{t+1}, and $\psi_{t+1}(y) = \int_S \Gamma(x, y)\psi_t(x)dx$. This last equality is just a version of the law of total probability: The probability of ending up at y is equal to the probability of going to y via x, weighted by the probability of being at x now, summed over all $x \in S$.

Now define map $\mathbf{M}\colon \mathscr{D} \to \mathscr{D}$, where $\mathscr{D} := \{\varphi \in L_1(S) \,|\, \varphi \geq 0 \text{ and } \int \varphi = 1\}$ is the space of densities on S, by

$$\mathbf{M}\colon \mathscr{D} \ni \psi \mapsto (\mathbf{M}\psi)(\cdot) := \int_S \Gamma(x, \cdot)\psi(x)dx \in \mathscr{D}. \tag{A.1}$$

With this definition our law of total probability rule for linking ψ_{t+1} and ψ_t can be written simply as $\psi_{t+1} = \mathbf{M}\psi_t$. Since the map $\mathbf{M}$ is deterministic, we have succeeded in transforming our stochastic system into a deterministic system to which standard methods of analysis may be applied. The only difficulty is that the state space is now $\mathscr{D}$ rather than S. The latter is finite dimensional, while the former clearly is not.

The map $\mathbf{M}$ is usually called the stochastic operator or Markov operator associated with Γ. There are many good expositions of Markov operators in economics, including Stokey, Lucas and Prescott (1989) and Futia (1982).

However those expositions treat the more general case, where $\Gamma(x, dy)$ does not necessarily have a density representation. Here it does, and it turns out that this extra structure is *very* useful for treating the models in this survey.

We wish to know when the difference equation $\psi_{t+1} = \mathbf{M}\psi_t$ has fixed points, and, more specifically, whether the system is globally stable in the sense that there is a unique fixed point ψ^*, and $\psi_t = \mathbf{M}^t\psi_0 \to \psi^*$ as $t \to \infty$ for all $\psi_0 \in \mathscr{D}$.[78] This is just ergodicity in the sense of Definition 3.1 on page 18.

Let $\|\cdot\|$ be the L_1 norm. Were $\mathbf{M}$ a uniform (Banach) contraction on $\mathscr{D}$, which is to say that $\exists\, \lambda < 1$ with $\|\mathbf{M}\psi - \mathbf{M}\psi'\| \leq \lambda\|\psi - \psi'\|$ for all $\psi, \psi' \in \mathscr{D}$, then ergodicity would hold because $\mathscr{D}$ is a closed subset of the complete metric space $L_1(S)$. Sadly, for continuous state Markov chains this uniform contraction property rarely holds. However it is often the case that $\mathbf{M}$ satisfies a weaker contraction condition:

Definition A.1 *Let $T\colon X \to X$, where (X, d) is a metric space. The map T is called a T2 contraction if $d(Tx, Tx') < d(x, x')$ for every $x \neq x'$ in X.*

T2 contractions maps distinct points strictly closer together. A sufficient condition for $\mathbf{M}\colon \mathscr{D} \to \mathscr{D}$ to satisfy the T2 property is given below. The essential requirement is communication across all regions of the state space. Although T2 contractions do not always have fixed points (examples in $\mathbb{R}$ are easy to construct), they do if the state space is compact! In fact if X is a compact set and $T\colon X \to X$ is a T2 contraction then T has unique fixed point $x^* \in X$ and $T^t x \to x^*$ as $t \to \infty$ for all $x \in X$. This is just what we require for ergodicity when $\mathbf{M}$ is thought of as a map on $\mathscr{D}$.

Now $\mathscr{D}$ is not itself a compact set in the L_1 norm topology, but it may be the case that every orbit $(\mathbf{M}^t\psi_0)_{t\geq 0}$ of $\mathbf{M}$ is compact when taken with its closure. (From now on, call a set with compact closure *precompact*). Such a property is called *Lagrange stability.*[79] And it turns out that Lagrange stability can substitute for compactness of the state space $\mathscr{D}$: If $\mathbf{M}$ is (a) a T2 contraction, and (b) Lagrange stable, then the associated Markov chain is ergodic.[80]

How to establish Lagrange stability? To check precompactness of orbits it

[78] Here $\mathbf{M}^t$ is t compositions of $\mathbf{M}$ with itself, and ψ_0 is the marginal distribution of x_0, so iterating the difference equation backwards gives $\psi_t = \mathbf{M}^t\psi_0$.
[79] That is, a self-mapping T on topological space X is called Lagrange stable if the set $\{T^t x \mid t \geq 0\}$ is precompact for every $x \in X$.
[80] The proof that Lagrange stability is sufficient is not hard. See Stachurski (2002, Theorem 5.2).

seems we must look at characterizations of compactness in L_1 (there is a famous one due to Kolmogorov), but Lasota (1994, Theorem 4.1) has proved that one need only check *weak* precompactness.[81] In fact it is sufficient to check weak precompactness of orbits starting from $\psi \in \mathscr{D}_0$, where $\mathscr{D}_0$ is a (norm) dense subset of $\mathscr{D}$. Weak compactness is much easier to work with than norm compactness. Several well-known conditions are available.

Using one such condition due to Dunford and Pettis, Mirman, Reffett and Stachurski (2004) show that Lasota's criterion for Lagrange stability is satisfied when (i) there exists a continuous "norm-like" function $V\colon S \to \mathbb{R}$ and constants $\alpha, \beta \in [0, \infty)$, $\alpha < 1$, such that

$$\int \Gamma(x, y)V(y)dy \leq \alpha V(x) + \beta, \quad \forall x \in S; \tag{A.2}$$

and (ii) there exists a continuous function $h\colon S \to \mathbb{R}$ such that $\sup_{x \in S} \Gamma(x, y) \leq h(y)$ for all $y \in S$. By V being norm-like is meant that V is nonnegative, and that the sets $\{x \in S : V(x) \leq a\}$ are precompact for all a. (For example, when $S = \mathbb{R}^n$ it is easy to convince yourself that $x \mapsto \|x\|$ is norm-like. Note that when S is a proper subset of $\mathbb{R}^n$ precompactness of sublevel sets refers to the *relative* Euclidean topology on S.)

Condition (i) is a standard drift condition, which pushes probability mass towards the center of the state space. This implies that orbits of the Markov process will be "tight." Tightness is a component of Dunford and Pettis' criterion for weak precompactness. Condition (ii) is just a technical condition which combines with (i) to fill out the requirements of the Dunford-Pettis criterion.

In the case of Proposition 3.1, we can take $S = (0, \infty)$, where $0 \notin S$ so that any stationary distribution we find is automatically nontrivial. One can then show that $V(x) = |\ln x|$ is norm-like on S, and a little bit of algebra shows that condition (i) holds for Γ given in (10). Also, one can show that (ii) holds when $h(y) := 1/y$.[82]

This takes care of Lagrange stability. Regarding T2 contractiveness, one can show that $\mathbf{M}$ is a T2 contraction whenever the set $\operatorname{supp} \mathbf{M}\psi \cap \operatorname{supp} \mathbf{M}\psi'$ has positive measure for all $\psi, \psi' \in \mathscr{D}$, where $\operatorname{supp} f := \{x \in S \,|\, f(x) \neq 0\}$. This

[81] Here is where the density structure is crucial. The operator $\mathbf{M}$ inherits nice properties from the fact that $\Gamma(x, dy)$ has a density representation. Also, we can work in L_1 rather than a space of measures. The former has a nice norm-dual space in L_∞—helpful when dealing with weak precompactness.

[82] For more details see Stachurski (2004).

basically says that probability mass is mixed across the state space—all areas of S communicate. In the case of (10) it is easy to show that $\operatorname{supp} \mathbf{M}\psi = (0, \infty) = S$ for every $\psi \in \mathscr{D}$. This is clearly sufficient for the condition.

A.2 Remaining proofs

The proof of Proposition 5.1 in Section 5.2 is now given. The first point is that the banks $b = 1, \ldots, B-1$ are equal-cost Bertrand competitors, and as a result always offer the interest rate r to all firms in equilibrium. The main issue is the optimal strategy of the last bank B. So consider the following strategy σ_B^* for B, which is illustrated with the help of Figure A.1. To firm n the bank offers i_n^* defined by $i_n^* = f[(n-1)/N] - 1$ if $n \leq \alpha_C N$. To the remaining firms B offers the interest rate r. (Without loss of generality, we suppose that the index of firms from 1 to N and the ranking of the offers made by B always coincide.) Let $\sigma^* \in \Sigma$ be the strategy where B offers σ_B^* and all other banks offer r.

For the strategy σ^* we have $\Omega(\sigma^*) = \{1\}$. The reason is that for $\alpha = n/N \leq \alpha_C$, firms $j = 1, \ldots, n+1$ all satisfy

$$\pi(n/N, m_j(\sigma^*)) \geq \pi(n/N, i_j^*) \geq \pi(n/N, i_{n+1}^*) = 0.$$

In which case $\alpha \notin \Omega(\sigma^*)$ by (20). Also, for $\alpha \in (\alpha_C, 1)$ we have $\pi(\alpha, m_n(\sigma^*)) \geq \pi(\alpha, r) \geq 0$ for all $n \in \{1, \ldots, N\}$, so again $\alpha \notin \Omega(\sigma^*)$. For the same reason, $1 \in \Omega(\sigma^*)$, because $\pi(1, m_n(\sigma^*)) \geq \pi(1, r) \geq 0$.

It follows that under this strategy $\alpha^{\text{pes}}(\sigma^*) = 1$. By (19) all firms enter. The profits of bank B are given by the sum of the regions P, Q and R, minus the region O, in Figure A.1. Here Q and $\bar{\alpha}$ are chosen so that $P+Q-O = 0$. Thus, $\bar{\alpha}$ is the break-even point for the bank, where it recoups all losses made by offering cheap loans to firms in the "critical mass" region $[0, \alpha_C]$. If $\ell/N \geq \bar{\alpha}$ and hence $R \geq 0$, the bank B makes positive profits.

It is not too hard to see that σ^* is indeed the optimal strategy in Σ for the banks. The banks $b = 1, \ldots, B-1$ always offer r. For B, strategy σ_B^* is optimal for the following reasons. First, if B offers interest rates to $n \in \{1, \ldots, N\}$ which are all less than or equal to those in σ_B^*, then all firms will enter as above, but B will make lower profits by (17). So suppose that B offers a schedule of rates $\{i_1^{**}, \ldots, i_N^{**}\}$ where $i_n^{**} > i_n^*$ for at least one n, and let k be the first such n. It is not difficult to see that the chain of logic whereby all firms enter now unravels: It must be that $k/N \leq \alpha_C$, because to other firms

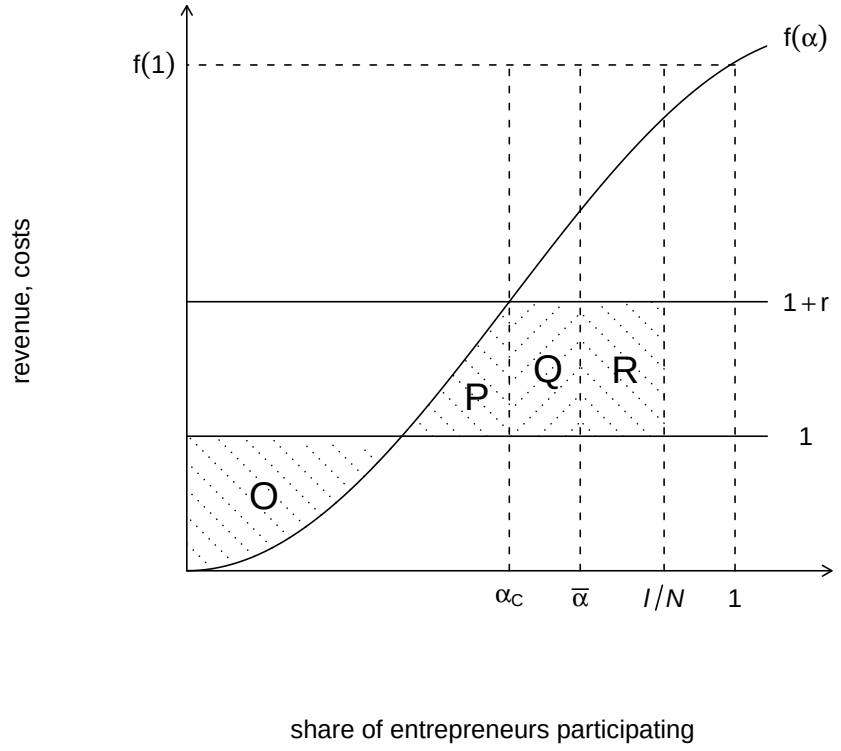


Fig. A.1. Coordination by the lead bank

B offers the rate r, which cannot be exceeded due to B's competitors. One can now check that $(k-1)/N \in \Omega(\sigma^{**})$, and in fact $(k-1)/N = \min \Omega(\sigma^{**})$. As a result, $\alpha^{\text{pes}}(\sigma^{**}) = (k-1)/N$, and precisely $k-1$ firms enter. Clearly the profits of B are lower for σ^{**} than for σ^{*}.

References

[1] Acemoglu, D. (1997): "Training and innovation in an imperfect labor market," *Review of Economic Studies,* 64 (3), 445–464.

[2] Acemoglu, D., S. Johnson and J. Robinson (this volume): "Institutions as the fundamental cause of long-run growth."

[3] Acemoglu, D. and F. Zilibotti (1997): "Was Prometheus unbound by chance? risk, diversification and growth," *Journal of Political Economy,* 105 (4), 709–751.

[4] Adsera, A. and D. Ray (1997): "History and coordination failure," *Journal of Economic Growth,* 3, 267–276.

[5] Aghion, P. and P. Bolton (1997): "A theory of trickle-down growth and development," *Review of Economic Studies,* 64, 151–172.

[6] Amir, R., L.J. Mirman and W.R. Perkins (1991): "One-sector nonclassical optimal growth: optimality conditions and comparative dynamics, *International Economic Review,* 32 (3), 625–644.

[7] Arthur, W. B. (1994): *Increasing Returns and Path Dependence in the Economy.* Ann Arbor: The University of Michigan Press.

[8] Azariadis, C. (1996): "The economics of poverty traps," *Journal of Economic Growth,* 1, 449–486.

[9] Azariadis, C. (2004): "The theory of poverty traps: what have we learned?" in *Poverty Traps,* (S. Bowles, S. Durlauf and K. Hoff, eds), in press.

[10] Azariadis, C. and A. Drazen (1990): "Threshold externalities in economic development," *Quarterly Journal of Economics,* 105, 501–526.

[11] Azariadis, C. and J. Stachurski (2004): "A forward projection of the cross-country income distribution," mimeo, Université Catholique de Louvain.

[12] Baker, E. (2004): "Institutional barriers to technology adoption in rural Africa," mimeo, Stanford University.

[13] Bandiera, O. and I. Rasul (2003): "Complementarities, social networks and technology adoption in northern Mozambique," mimeo.

[14] Banerjee, A. (2003): "The two poverties," mimeo, Massachusetts Institute of Technology.

[15] Banerjee, A. and A. Newman (1993): "Occupational choice and the process of development," *Journal of Political Economy,* 101, 274–298.

[16] Bardhan, P. (1997): "Corruption and development: a review of the issues," *Journal of Economic Literature,* 35, 1320–1346.

[17] Barrett, C. B., M. Bezuneh and A. Aboud (2001): "Income diversification, poverty traps and policy shocks in Côte d'Ivoire and Kenya," *Food Policy*, in press.

[18] Barrett, C. B. and B. M. Swallow (2003): "Fractal poverty traps," Strategies and Analysis for Growth and Access Working Paper, Cornell and Clark Atlanta Universities.

[19] Basu, K. and P. H. Van (1998): "The economics of child labor," *American Economic Review,* 88 (3), 412–427.

[20] Beath, J., Y. Katsoulacos and D. Ulph (1995): "Game-theoretic approaches to the modeling of technological change," in *Handbook of the Economics of Innovation and Technological Change,* (P. Stoneman, ed) Blackwell.

[21] Bianchi, M. (1997): "Testing for convergence: evidence from nonparametric multimodality tests," *Journal of Applied Econometrics,* 12, 393–409.

[22] Bloom, D. E., D. Canning and J. Sevilla (2003): "Geography and poverty traps," *Journal of Economic Growth,* 8, 355–378.

[23] Bowles, S., S. N. Durlauf and K. Hoff, eds. (2004): *Poverty Traps,* in press.

[24] Brock, W. A. and L. Mirman (1972): "Optimal economic growth and uncertainty: the discounted case," *Journal of Economic Theory,* 4, 479–513.

[25] Cabarello, R. and R. K. Lyons (2002): "The case for external economies," in *Political Economy, Growth and Business Cycles* (A. Cukierman et al., eds.) MIT Press, Cambridge.

[26] Cass, D. (1965): "Optimum growth in an aggregative model of capital accumulation," *Review of Economic Studies,* 32, 233–240.

[27] Chen, S. and M. Ravallion (2001): "How did the world's poor fare in the 1990s?" *Review of Income and Wealth,* 47 (3), 238–300.

[28] Checchi, D. and C. García-Peñalosa (2004): "Risk and the distribution of human capital," *Economics Letters,* 82 (1), 53–61.

[29] Conley, T. G. and C. R. Udry (2003): "Learning about a new technology: pinapple in Ghana," mimeo

[30] Da Rin, M. and T. Hellmann (2002): "Banks as catalysts for industrialization," *Journal of Financial Intermediation,* 11, 366–397.

[31] Dasgupta, P. (2003): "World poverty: causes and pathways," mimeo, University of Cambridge.

[32] Dasgupta, P. and D. Ray (1986): "Inequality as a determinant of malnutrition and unemployment," *The Economic Journal,* 96, 1011–1034.

[33] David, P. A. (1994) "Why are institutions the 'carriers of history'?: Path dependence and the evolution of conventions, organizations and institutions," *Structural Change and Economic Dynamics,* 5 (2), 205–220

[34] Dechert, W. D. and K. Nishimura (1983): "A complete characterization of optimal growth paths in an aggregated model with non-concave production function, *Journal of Economic Theory,* 31, 332–354.

[35] Den Haan, W. J. (1995): "Convergence of stochastic growth models: the importance of understanding why income levels differ," *Journal of Monetary Economics,* 35, 65–82.

[36] Dercon, S. (1998): "Wealth, risk and activity choice: cattle in western Tanzania," *Journal of Development Economics,* 55, 1–42.

[37] Dercon, S. (2003) "Risk and poverty: a selective review of the issues," mimeo, Oxford University.

[38] Desdoigts, A. (1999): "Patters of economic development and the formation of clubs," *Journal of Economic Growth,* 4 (3), 305–330.

[39] De Soto, H. (1989): *The Other Path: The Invisible Revolution in the Third World,* New York: Harper and Row.

[40] Dimaria, C. H. and C. Le Van (2002): "Optimal growth, debt, corruption and R&D," *Macroeconomic Dynamics,* 6, 597–613.

[41] Durlauf, S. N. (1993): "Nonergodic economic growth," *Review of Economic Studies,* 60, 349–366.

[42] Durlauf, S. N. and P. A. Johnson (1995): "Multiple regimes and cross-country growth behavior", *Journal of Applied Econometrics,* 10 (4), 365-384.

[43] Durlauf, S.N. (2004): "Neighborhood effects," in *Handbook of Urban and Regional Economics,* (J.V. Henderson and J.F. Thisse, eds), in press.

[44] Easterly, W. (2001): *The Elusive Quest for Qrowth: Economists' Adventures and Misadventures in the Tropics,* The MIT Press, Cambridge, Massachusetts.

[45] Easterly, W., M. Kremer, L. Pritchett and L. H. Summers (1993): "Good policy or good luck: country growth performance and temporary shocks," *Journal of Monetary Economics,* 32, 459–484.

[46] Easterly, W. and R. Levine (2000): "It's not factor accumulation: stylized facts and growth models," *World Bank Economic Review,* 15 (2), 177–219.

[47] Engerman, S. and K. L. Sokoloff (2004): "The persistence of poverty in the Americas: the role of institutions," in *Poverty Traps,* (S. Bowles, S. Durlauf and K. Hoff, eds), in press.

[48] Feyrer, J. (2003): "Convergence by parts," mimeo, Dartmouth College.

[49] Freeman, S. (1996): "Equilibrium income inequality among identical agents," *Journal of Political Economy,* 104 (5), 1047–1064.

[50] Futia, C. A. (1982): "Invariant distributions and the limiting behavior of Markovian economic models," *Econometrica,* 50, 377–408.

[51] Galor, O. and J. Zeira (1993): "Income distribution and macroeconomics," *Review of Economic Studies,* 60, 35–52.

[52] Glynn, P. W. and S. G. Henderson (2001): "Computing densities for Markov chains via simulation," *Mathematics of Operations Research,* 26, 375–400.

[53] Gradstein, M. (2004): "Governance and growth," *Journal of Development Economics,* 73, 505–518.

[54] Graham, B. S. and J. Temple (2004): "Rich nations poor nations: how much can multiple equilibria explain?" mimeo (revised version of CEPR Discussion Paper 3046).

[55] Greif, A., P. Milgrom and B.R. Weingast (1994): "Coordination, commitment and enforcement: the case of merchant guilds," *Journal of Political Economy,* 102 (4), 745–776.

[56] Hall, R. E. and C. I. Jones (1999): "Why do some countries produce so much more output per worker than others?" *Quarterly Journal of Economics,* 114 (1), 84–116.

[57] Heston, A., R. Summers and B. Aten (2002): "Penn World Table version 6.1." Center for International Comparisons, University of Pennsylvania.

[58] Hoff, K. (2000): "Beyond Rosenstein-Rodan: the modern theory of coordination problems in development," *Proceedings of the World Bank Annual Conference on Development Economics 2000*, in press.

[59] Hoff, K. and A. Sen (2004): "The kin system as a poverty trap," in *Poverty Traps,* (S. Bowles, S. Durlauf and K. Hoff, eds), in press.

[60] Hoff, K. and Pandey (2004): "Belief systems and durable inequalities: an experimental investigation of Indian caste," World Bank Policy Research Working Paper 3351.

[61] Hoff, K. and J. Stiglitz (2004): "After the Big Bang? Obstacles to the emergence of the rule of law in post-communist societies," *American Economic Review,* 94 (3) 753–763.

[62] Ihaka, R. and R. Gentleman (1996): "R: a language for data analysis and graphics," *Journal of Computational and Graphical Statistics,* 5 (3), 299–314.

[63] Jalan, J. and M. Ravallion (2002): "Geographic poverty traps? A micro model of consumption growth in rural China," *Journal of Applied Econometrics,* 17, 329–346.

[64] Johnson, P. A. (2004): "A continuous state space approach to convergence by parts," mimeo, Vassar College.

[65] Kehoe, T. and D. Levine (1993): "Debt constrained asset markets," *Review of Economic Studies,* 60, 865–888.

[66] King, R. G. and T. Rebelo (1993): "Transitional dynamics and economic growth in the neoclassical model," *American Economic Review,* 83 (4), 908–931.

[67] Kiyotaki, N. and J. H. Moore (1997): "Credit cycles," *Journal of Political Economy,* 105 (2), 211–248.

[68] Krugman, P. (1991): "History versus expectations," *The Quarterly Journal of Economics,* 106 (2), 651–667.

[69] Koopmans, T. (1965): "On the concept of optimal economic growth," *Pontificae Academiae Scientiarum Scripta Varia,* 28, 225–300.

[70] Kremer, M. (1993): "The O-ring theory of economic development," *Quarterly Journal of Economics,* August, 551–575.

[71] Lambsdorff, J. G. (2003): "How corruption affects persistent capital flows," *Economics of Governance,* 4 (3), 229–243.

[72] Lasota, A. (1994), "Invariant principle for discrete time dynamical systems," *Universitatis Iagellonicae Acta Mathematica* 31, 111–127.

[73] Limao, N. and A.J. Venables (2001): "Infrastructure, geographical disadvantage, transport costs and trade," *World Bank Economic Review,* 15, 451–479.

[74] Ljungqvist, L. (1993): "Economic underdevelopment: the case of the missing market for human capital," *Journal of Development Economics,* 40, 219–239.

[75] Loury, G. C. (1981): "Intergenerational transfers and the distribution of earnings," *Econometrica,* 49, 843–867.

[76] Lucas, R. E. Jr, (1986): "Adaptive behavior and economic theory," *Journal of Business,* 59 (4), 401–426.

[77] Lucas, R. E. (1990): "Why doesn't capital flow from rich to poor countries?" *American Economic Review,* 80 (2), 92-96

[78] Maddison, A. (1995): *Monitoring the World Economy,* OECD Development Center, Paris.

[79] Magill, M. and K. Nishimura (1984): "Impatience and accumulation," *Journal of Mathematical Analysis and Applications,* 98, 270–281.

[80] Majumdar, M., T. Mitra and Y. Nyarko (1989): "Dynamic optimization under uncertainty: non-convex feasible set" in *Joan Robinson and Modern Economic Theory,* (G. R. Feiwel, ed) MacMillan Press, New York.

[81] Matsuyama, K. (1991): "Increasing returns, industrialization and indeterminacy of equilibrium," *Quarterly Journal of Economics,* 106, 617–650.

[82] Matsuyama, K. (1995): "Complementarities and cumulative processes in models of monopolistic competition," *Journal of Economic Literature,* 33, 701–729.

[83] Matsuyama, K. (1997): "Complementarity, instability and multiplicity," *Japanese Economic Review,* 48 (3), 240–266.

[84] Matsuyama, K. (2000): "Endogenous inequality," *Review of Economic Studies,* 67, 743–759.

[85] Matsuyama, K. (2004): "Financial market globalization, symmetry-breaking, and endogenous inequality of nations," *Econometrica,* 72, 853–884.

[86] Matsuyama, K. and A. Ciccone (1996): "Start-up costs and pecuniary externalities as barriers to economic development," *Journal of Development Economics,* 49, 33–59.

[87] Mirman, L.J., O.F. Morand and K. Reffett (2004), "A qualitative approach to Markovian equilibrium in infinite horizon economies with capital," mimeo, Arizona State University.

[88] Mirman, L. J., K. Reffett and J. Stachurski (2004): "Some stability results for Markovian economic semigroups," *International Journal of Economic Theory,* in press.

[89] Mokyr, J. (2002): *The Gifts of Athena: Historical Origins of the Knowledge Economy,* Princeton University Press, Princeton New Jersey.

[90] Mookherjee, D. and D. Ray (2001) *Readings in the Theory of Economic Development,* Blackwell Publishing, New York.

[91] Mookherjee, D. and D. Ray (2003): "Persistent inequality," *Review of Economic Studies,* 70, 369–393.

[92] Morduch, J. (1990): "Risk, production and saving: Theory and evidence from Indian households," Harvard University, mimeo.

[93] Murphy, K. M., A. Shleifer and R. W. Vishny (1989): "Industrialization and the big push," *Journal of Political Economy,* 97, 1003–1026.

[94] Murphy, K. M., A. Shleifer and R. W. Vishny (1993): "Why is rent-seeking so costly to growth?" *American Economic Review,* 83 (2), 409–414.

[95] Nelson, R. R. (1956): "A theory of the low level equilibrium trap", *American Economic Review,* 46, 894–908.

[96] Nishimura, K. and J. Stachurski (2004): "Stability of stochastic optimal growth models: a new approach," *Journal of Economic Theory,* in press.

[97] North (1993): "The new institutional economics and development," WUSTL Economics Working Paper Archive.

[98] North (1995): "Some fundamental puzzles in economic history/development," WUSTL Economics Working Paper Archive.

[99] Nurkse, R (1953): *Problems of Capital-Formation in Underdeveloped Countries,* 1962 edition, New York: Oxford Univeristy Press.

[100] Ottaviano, A. and J.F. Thisse (2004): "Agglomeration and economic geography," in *Handbook of Urban and Regional Economics,* (J.V. Henderson and J.F. Thisse, eds), in press.

[101] Parente, S. L. and E. C. Prescott (this volume): "A unified theory of the evolution of international income levels."

[102] Piketty, T. (1997): "The dynamics of the wealth distribution and the interest rate with credit rationing," *Review of Economic Studies,* 64, 173–189.

[103] Platteau, J-P. (2000): *Institutions, Social Norms and Economic Development* Amsterdam: Harwood Publishers

[104] Prescott, E. C. (1998): "Needed: a theory of total factor productivity," *International Economic Review,* 39, 529–549.

[105] Pritchett, L. (1997): "Divergence, big time," *Journal of Economic Perspectives,* 11 (3), 3–17.

[106] Quah, D. T. (1993): "Empirical cross-section dynamics in economic growth," *European Economic Review,* 37, 426-434.

[107] Quah, D. (1996): "Convergence empirics across economies with (some) capital mobility," *Journal of Economic Growth,* 1, 95–124.

[108] Radlet, S. (2004): "Aid effectiveness and the Millenium Development Goals," manuscript prepared for the Millenium Project Task Force, United Nations Development Group.

[109] Ray, D. (1990): "Income distribution and macroeconomic behavior," mimeo, New York University.

[110] Ray, D. (2003): "Aspirations, poverty and economic change," mimeo, New York University.

[111] Ray, D. and P. Streufert (1993): "Dynamic equilibria with unemployment due to undernourishment," *Economic Theory,* 3, 61–85.

[112] Redding, S. and A. J. Venables (2004): "Economic geography and international inequality," *Journal of International Economics,* 62, 53–82.

[113] Rodríguez-Clare, A. (1996): "The division of labor and economic development," *Journal of Development Economics,* 49, 3–32.

[114] Rodrik, D. (1996): "Coordination failures and government policy: A model with applications to East Asia and Eastern Europe," *Journal of International Economics,* 40, 1–22.

[115] Romer, P. M. (1986): "Increasing returns and long-run growth," *Journal of Political Economy,* 94 (5), 1002-1037.

[116] Romer, P. M. (1990): "Are nonconvexities important for understanding growth?," *American Economic Review,* 80 (2), 97–103.

[117] Rosenstein-Rodan, P. (1943): "The problem of industrialization of eastern and south-eastern Europe," *Economic Journal,* 53, 202–211

[118] Rostow, W. W. (1975): *How it All Began: Origins of the Modern Economy,* McGraw-Hill, New York.

[119] Rostow, W. W. (1990): *The Stages of Economic Growth: A Noncommunist Manifesto,* 3rd edition, Cambridge University Press, Cambridge.

[120] Sachs, J. D., J. W. McArthur, G. Schmidt-Traub, M. Kruk, C. Bahadur, M. Faye and G. McCord, "Ending Africa's poverty trap," mimeo

[121] Simon, H. A. (1986): "Rationality in psychology and economics," *Journal of Business,* 59 (4), 209–224.

[122] Skiba, A. K. (1978): "Optimal growth with a convex-concave production function," *Econometrica,* 46, 527–539.

[123] Solow, R. M. (1956): "A contribution to the theory of economic growth." *Quarterly Journal of Economics,* 70, 65–94.

[124] Stachurski, J. (2002): "Stochastic optimal growth with unbounded shock," *Journal of Economic Theory,* 106, 45–60.

[125] Stachurski, J. (2004): "Stochastic economic dynamics," mimeo, the University of Melbourne.

[126] Starrett, D. (1978): "Market allocations of location choice in a model with free mobility," *Journal of Economic Theory,* 17, 21–37.

[127] Stokey, N. L., R. E. Lucas and E. C. Prescott (1989): *Recursive Methods in Economic Dynamics,* Harvard University Press, Massachusetts.

[128] Sugden, R. (1989): "Spontaneous order," *Journal of Economic Perspectives,* 3 (4), 85–97.

[129] Tirole, J. (1996): "A theory of collective reputations (with applications to persistence of corruption and to firm quality)," *Review of Economic Studies,* 63, 1–22.

[130] Tsiddon, D. (1992): "A moral hazard trap to growth," *International Economic Review,* 33 (2), 299–321.

[131] Van Huyck, J. B., J. P. Cook and R. C. Battalio (1997): "Adaptive behavior and coordination failure," *Journal of Economic Behavior and Organization,* 32 (4), 483–503

[132] Young, A. A. (1928): "Increasing returns and economic progress," *Economic Journal,* 28, 527–542.

[133] Zilibotti, F. (1995): "A Rostovian model of endogenous growth and underdevelopment traps," *European Economic Review,* 39, 1569–1602.

Institutions as the Fundamental Cause of Long-Run Growth*

Daron Acemoglu
Department of Economics, MIT

Simon Johnson
Sloan School of Management, MIT

James Robinson
Departments of Political Science and Economics, Berkeley

April 29, 2004

*Prepared for the *Handbook of Economic Growth* edited by Philippe Aghion and Steve Durlauf. We thank the editors for their patience and Leopoldo Fergusson, Pablo Querubín and Barry Weingast for their helpful suggestions.

Abstract

This paper develops the empirical and theoretical case that differences in economic institutions are the fundamental cause of differences in economic development. We first document the empirical importance of institutions by focusing on two "quasi-natural experiments" in history, the division of Korea into two parts with very different economic institutions and the colonization of much of the world by European powers starting in the fifteenth century. We then develop the basic outline of a framework for thinking about why economic institutions differ across countries. Economic institutions determine the incentives of and the constraints on economic actors, and shape economic outcomes. As such, they are social decisions, chosen for their consequences. Because different groups and individuals typically benefit from different economic institutions, there is generally a conflict over these social choices, ultimately resolved in favor of groups with greater political power. The distribution of political power in society is in turn determined by political institutions and the distribution of resources. Political institutions allocate *de jure* political power, while groups with greater economic might typically possess greater *de facto* political power. We therefore view the appropriate theoretical framework as a dynamic one with political institutions and the distribution of resources as the state variables. These variables themselves change over time because prevailing economic institutions affect the distribution of resources, and because groups with de facto political power today strive to change political institutions in order to increase their de jure political power in the future. Economic institutions encouraging economic growth emerge when political institutions allocate power to groups with interests in broad-based property rights enforcement, when they create effective constraints on power-holders, and when there are relatively few rents to be captured by power-holders. We illustrate the assumptions, the workings and the implications of this framework using a number of historical examples.

1 Introduction

1.1 The question

The most trite yet crucial question in the field of economic growth and development is: Why are some countries much poorer than others? Traditional neoclassical growth models, following Solow (1956), Cass (1965) and Koopmans (1965), explain differences in income per capita in terms of different paths of factor accumulation. In these models, cross-country differences in factor accumulation are due either to differences in saving rates (Solow), preferences (Cass-Koopmans), or other exogenous parameters, such as total factor productivity growth. More recent incarnations of growth theory, following Romer (1986) and Lucas (1988), endogenize steady-state growth and technical progress, but their explanation for income differences is similar to that of the older theories. For instance, in the model of Romer (1990), a country may be more prosperous than another if it allocates more resources to innovation, but what determines this is essentially preferences and properties of the technology for creating 'ideas'.[1]

Though this theoretical tradition is still vibrant in economics and has provided many insights about the mechanics of economic growth, it has for a long time seemed unable to provide a *fundamental* explanation for economic growth. As North and Thomas (1973, p. 2) put it: "the factors we have listed (innovation, economies of scale, education, capital accumulation etc.) are not causes of growth; they *are* growth" (italics in original). Factor accumulation and innovation are only *proximate* causes of growth. In North and Thomas's view, the fundamental explanation of comparative growth is differences in *institutions*.

What are institutions exactly? North (1990, p. 3) offers the following definition: "Institutions are the rules of the game in a society or, more formally, are the humanly devised constraints that shape human interaction." He goes on to emphasize the key implications of institutions since, "In consequence they structure incentives in human exchange, whether political, social, or economic."

Of primary importance to economic outcomes are the *economic institutions* in society such as the structure of property rights and the presence and perfection of markets. Economic institutions are important because they influence the structure of economic

[1] Although some recent contributions to growth theory emphasize the importance of economic policies, such as taxes, subsidies to research, barriers to technology adoption and human capital policy, they typically do not present an explanation for why there are differences in these policies across countries.

incentives in society. Without property rights, individuals will not have the incentive to invest in physical or human capital or adopt more efficient technologies. Economic institutions are also important because they help to allocate resources to their most efficient uses, they determine who gets profits, revenues and residual rights of control. When markets are missing or ignored (as they were in the Soviet Union, for example), gains from trade go unexploited and resources are misallocated. Societies with economic institutions that facilitate and encourage factor accumulation, innovation and the efficient allocation of resources will prosper.

Central to this chapter and to much of political economy research on institutions is that economic institutions, and institutions more broadly, are *endogenous*; they are, at least in part, determined by society, or a segment of it. Consequently, the question of why some societies are much poorer than others is closely related to the question of why some societies have much "worse economic institutions" than others.

Even though many scholars including John Locke, Adam Smith, John Stuart Mill, Douglass North and Robert Thomas have emphasized the importance of economic institutions, we are far from a useful framework for thinking about how economic institutions are determined and why they vary across countries. In other words, while we have good reason to believe that economic institutions matter for economic growth, we lack the crucial *comparative static* results which will allow us to explain why equilibrium economic institutions differ (and perhaps this is part of the reason why much of the economics literature has focused on the proximate causes of economic growth, largely neglecting fundamental institutional causes).

This chapter has three aims. First, we selectively review the evidence that differences in economic institutions are a fundamental cause of cross-country differences in prosperity. Second, we outline a framework for thinking about why economic institutions vary across countries. We emphasize the potential comparative static results of this framework and also illustrate the key mechanisms through a series of historical examples and case studies. Finally, we highlight a large number of areas where future theoretical and empirical work would be very fruitful.

1.2 The Argument

The basic argument of this chapter can be summarized as follows:

1. Economic institutions matter for economic growth because they shape the incentives of key economic actors in society, in particular, they influence investments in physical and human capital and technology, and the organization of production. Al-

though cultural and geographical factors may also matter for economic performance, differences in economic institutions are the major source of cross-country differences in economic growth and prosperity. Economic institutions not only determine the aggregate economic growth potential of the economy, but also an array of economic outcomes, including the distribution of resources in the future (i.e., the distribution of wealth, of physical capital or human capital). In other words, they influence not only the size of the aggregate pie, but how this pie is divided among different groups and individuals in society. We summarize these ideas schematically as (where the subscript t refers to current period and $t+1$ to the future):

$$\text{economic institutions}_t \Longrightarrow \begin{cases} \text{economic performance}_t \\ \text{distribution of resources}_{t+1} \end{cases}.$$

2. Economic institutions are endogenous. They are determined as collective choices of the society, in large part for their economic consequences. However, there is no guarantee that all individuals and groups will prefer the same set of economic institutions because, as noted above, different economic institutions lead to different distributions of resources. Consequently, there will typically be a *conflict of interest* among various groups and individuals over the choice of economic institutions. So how are equilibrium economic institutions determined? If there are, for example, two groups with opposing preferences over the set of economic institutions, which group's preferences will prevail? The answer depends on the *political power* of the two groups. Although the efficiency of one set of economic institutions compared with another may play a role in this choice, political power will be the ultimate arbiter. Whichever group has more political power is likely to secure the set of economic institutions that it prefers. This leads to the second building block of our framework:

$$\text{political power}_t \Longrightarrow \text{economic institutions}_t$$

3. Implicit in the notion that political power determines economic institutions is the idea that there are conflicting interests over the distribution of resources and therefore indirectly over the set of economic institutions. But why do the groups with conflicting interests not agree on the set of economic institutions that maximize aggregate growth (the size of the aggregate pie) and then use their political power simply to determine the distribution of the gains? Why does the exercise of political power lead to economic inefficiencies and even poverty? We will explain that this is because there are commitment problems inherent in the use of political power. Individuals who have political power

cannot commit not to use it in their best interests, and this commitment problem creates an inseparability between efficiency and distribution because credible compensating transfers and side-payments cannot be made to offset the distributional consequences of any particular set of economic institutions.

4. The distribution of political power in society is also endogenous, however. In our framework, it is useful to distinguish between two components of political power, which we refer to as *de jure (institutional)* and *de facto political power*. Here de jure political power refers to power that originates from the *political institutions* in society. Political institutions, similarly to economic institutions, determine the constraints on and the incentives of the key actors, but this time in the political sphere. Examples of political institutions include the form of government, for example, democracy vs. dictatorship or autocracy, and the extent of constraints on politicians and political elites. For example, in a monarchy, political institutions allocate all de jure political power to the monarch, and place few constraints on its exercise. A constitutional monarchy, in contrast, corresponds to a set of political institutions that reallocates some of the political power of the monarch to a parliament, thus effectively constraining the political power of the monarch. This discussion therefore implies that:

$$\text{political institutions}_t \Longrightarrow \text{ de jure political power}_t$$

5. There is more to political power than political institutions, however. A group of individuals, even if they are not allocated power by political institutions, for example as specified in the constitution, may nonetheless possess political power. Namely, they can revolt, use arms, hire mercenaries, co-opt the military, or use economically costly but largely peaceful protests in order to impose their wishes on society. We refer to this type of political power as de facto political power, which itself has two sources. First, it depends on the ability of the group in question to solve its collective action problem, i.e., to ensure that people act together, even when any individual may have an incentive to free ride. For example, peasants in the Middle Ages, who were given no political power by the constitution, could sometimes solve the collective action problem and undertake a revolt against the authorities. Second, the de facto power of a group depends on its economic resources, which determine both their ability to use (or misuse) existing political institutions and also their option to hire and use force against different groups. Since we do not yet have a satisfactory theory of when groups are able to solve their collective action problems, our focus will be on the second source of de facto political

power, hence:

$$\text{distribution of resources}_t \Longrightarrow \text{de facto political power}_t$$

6. This brings us to the evolution of one of the two main *state variables* in our framework, political institutions (the other state variable is the distribution of resources, including distribution of physical and human capital stocks etc.). Political institutions and the distribution of resources are the state variables in this dynamic system because they typically change relatively slowly, and more importantly, they determine economic institutions and economic performance both directly and indirectly. Their direct effect is straightforward to understand. If political institutions place all political power in the hands of a single individual or a small group, economic institutions that provide protection of property rights and equal opportunity for the rest of the population are difficult to sustain. The indirect effect works through the channels discussed above: political institutions determine the distribution of de jure political power, which in turn affects the choice of economic institutions. This framework therefore introduces a natural concept of *a hierarchy of institutions*, with political institutions influencing equilibrium economic institutions, which then determine economic outcomes.

Political institutions, though slow changing, are also endogenous. Societies transition from dictatorship to democracy, and change their constitutions to modify the constraints on power holders. Since, like economic institutions, political institutions are collective choices, the distribution of political power in society is the key determinant of their evolution. This creates a tendency for persistence: political institutions allocate de jure political power, and those who hold political power influence the evolution of political institutions, and they will generally opt to maintain the political institutions that give them political power. However, de facto political power occasionally creates changes in political institutions. While these changes are sometimes discontinuous, for example when an imbalance of power leads to a revolution or the threat of revolution leads to major reforms in political institutions, often they simply influence the way existing political institutions function, for example, whether the rules laid down in a particular constitution are respected as in most functioning democracies, or ignored as in current-day Zimbabwe. Summarizing this discussion, we have:

$$\text{political power}_t \Longrightarrow \text{political institutions}_{t+1}$$

Putting all these pieces together, a schematic (and simplistic) representation of our

framework is as follows:

$$\left.\begin{array}{l} \text{political} \\ \text{institutions}_t \end{array} \Longrightarrow \begin{array}{c} \text{de jure} \\ \text{political} \\ \text{power}_t \end{array}\right\} \Longrightarrow \text{economic institutions}_t \Longrightarrow \left\{\begin{array}{c} \text{economic} \\ \text{performance}_t \\ \& \\ \text{distribution} \\ \text{of resources}_{t+1} \end{array}\right.$$

$$\left.\begin{array}{l} \text{distribution} \\ \text{of resources}_t \end{array} \Longrightarrow \begin{array}{c} \text{de facto} \\ \text{political} \\ \text{power}_t \end{array}\right\} \Longrightarrow \text{political institutions}_{t+1}$$

The two state variables are political institutions and the distribution of resources, and the knowledge of these two variables at time t is sufficient to determine all the other variables in the system. While political institutions determine the distribution of de jure political power in society, the distribution of resources influences the distribution of de facto political power at time t. These two sources of political power, in turn, affect the choice of economic institutions and influence the future evolution of political institutions. Economic institutions determine economic outcomes, including the aggregate growth rate of the economy and the distribution of resources at time $t+1$. Although economic institutions are the essential factor shaping economic outcomes, they are themselves endogenous and determined by political institutions and distribution of resources in society.

There are two sources of persistence in the behavior of the system: first, political institutions are durable, and typically, a sufficiently large change in the distribution of political power is necessary to cause a change in political institutions, such as a transition from dictatorship to democracy. Second, when a particular group is rich relative to others, this will increase its de facto political power and enable it to push for economic and political institutions favorable to its interests. This will tend to reproduce the initial relative wealth disparity in the future. Despite these tendencies for persistence, the framework also emphasizes the potential for change. In particular, "shocks", including changes in technologies and the international environment, that modify the balance of (de facto) political power in society and can lead to major changes in political institutions and therefore in economic institutions and economic growth.

A brief example might be useful to clarify these notions before commenting on some of the underlying assumptions and discussing comparative statics. Consider the development of property rights in Europe during the Middle Ages. There is no doubt that lack of property rights for landowners, merchants and proto- industrialists was detrimental to economic growth during this epoch. Since political institutions at the time placed political power in the hands of kings and various types of hereditary monarchies,

such rights were largely decided by these monarchs. Unfortunately for economic growth, while monarchs had every incentive to protect their own property rights, they did not generally enforce the property rights of others. On the contrary, monarchs often used their powers to expropriate producers, impose arbitrary taxation, renege on their debts, and allocate the productive resources of society to their allies in return for economic benefits or political support. Consequently, economic institutions during the Middle Ages provided little incentive to invest in land, physical or human capital, or technology, and failed to foster economic growth. These economic institutions also ensured that the monarchs controlled a large fraction of the economic resources in society, solidifying their political power and ensuring the continuation of the political regime.

The seventeenth century, however, witnessed major changes in the economic and political institutions that paved the way for the development of property rights and limits on monarchs' power, especially in England after the Civil War of 1642 and the Glorious Revolution of 1688, and in the Netherlands after the Dutch Revolt against the Hapsburgs. How did these major institutional changes take place? In England, for example, until the sixteenth century the king also possessed a substantial amount of de facto political power, and leaving aside civil wars related to royal succession, no other social group could amass sufficient de facto political power to challenge the king. But changes in the English land market (Tawney, 1941) and the expansion of Atlantic trade in the sixteenth and seventeenth centuries (Acemoglu, Johnson and Robinson, 2002b) gradually increased the economic fortunes, and consequently the de facto power of landowners and merchants. These groups were diverse, but contained important elements that perceived themselves as having interests in conflict with those of the king: while the English kings were interested in predating against society to increase their tax incomes, the gentry and merchants were interested in strengthening their property rights.

By the seventeenth century, the growing prosperity of the merchants and the gentry, based both on internal and overseas, especially Atlantic, trade, enabled them to field military forces capable of defeating the king. This de facto power overcame the Stuart monarchs in the Civil War and Glorious Revolution, and led to a change in political institutions that stripped the king of much of his previous power over policy. These changes in the distribution of political power led to major changes in economic institutions, strengthening the property rights of both land and capital owners and spurred a process of financial and commercial expansion. The consequence was rapid economic growth, culminating in the Industrial Revolution, and a very different distribution of

economic resources from that in the Middle Ages.

It is worth returning at this point to two critical assumptions in our framework. First, why do the groups with conflicting interests not agree on the set of economic institutions that maximize aggregate growth? So in the case of the conflict between the monarchy and the merchants, why does the monarchy not set up secure property rights to encourage economic growth and tax some of the benefits? Second, why do groups with political power want to change political institutions in their favor? For instance, in the context of the example above, why did the gentry and merchants use their de facto political power to change political institutions rather than simply implement the policies they wanted? The answers to both questions revolve around issues of *commitment* and go to the heart of our framework.

The distribution of resources in society is an inherently conflictual, and therefore political, decision. As mentioned above, this leads to major commitment problems, since groups with political power cannot commit to not using their power to change the distribution of resources in their favor. For example, economic institutions that increased the security of property rights for land and capital owners during the Middle Ages would not have been credible as long as the monarch monopolized political power. He could promise to respect property rights, but then at some point, renege on his promise, as exemplified by the numerous financial defaults by medieval kings (e.g., Veitch, 1986). Credible secure property rights necessitated a reduction in the political power of the monarch. Although these more secure property rights would foster economic growth, they were not appealing to the monarchs who would lose their rents from predation and expropriation as well as various other privileges associated with their monopoly of political power. This is why the institutional changes in England as a result of the Glorious Revolution were not simply conceded by the Stuart kings. James II had to be deposed for the changes to take place.

The reason why political power is often used to change political institutions is related. In a dynamic world, individuals care not only about economic outcomes today but also in the future. In the example above, the gentry and merchants were interested in their profits and therefore in the security of their property rights, not only in the present but also in the future. Therefore, they would have liked to use their (de facto) political power to secure benefits in the future as well as the present. However, commitment to future allocations (or economic institutions) was not possible because decisions in the future would be decided by those who had political power in the future with little reference to past promises. If the gentry and merchants would have been sure to maintain their de

facto political power, this would not have been a problem. However, de facto political power is often transient, for example because the collective action problems that are solved to amass this power are likely to resurface in the future, or other groups, especially those controlling de jure power, can become stronger in the future. Therefore, any change in policies and economic institutions that relies purely on de facto political power is likely to be reversed in the future. In addition, many revolutions are followed by conflict within the revolutionaries. Recognizing this, the English gentry and merchants strove not just to change economic institutions in their favor following their victories against the Stuart monarchy, but also to alter political institutions and the future allocation of de jure power. Using political power to change political institutions then emerges as a useful strategy to make gains more durable. The framework that we propose, therefore, emphasizes the importance of political institutions, and changes in political institutions, as a way of manipulating future political power, and thus indirectly shaping future, as well as present, economic institutions and outcomes.

This framework, though abstract and highly simple, enables us to provide some preliminary answers to our main question: why do some societies choose "good economic institutions"? At this point, we need to be more specific about what good economic institutions are. A danger we would like to avoid is that we define good economic institutions as those that generate economic growth, potentially leading to a tautology. This danger arises because a given set of economic institutions may be relatively good during some periods and bad during others. For example, a set of economic institutions that protects the property rights of a small elite might not be inimical to economic growth when all major investment opportunities are in the hands of this elite, but could be very harmful when investments and participation by other groups are important for economic growth (see Acemoglu, 2003b). To avoid such a tautology and to simplify and focus the discussion, throughout we think of good economic institutions as those that provide security of property rights and relatively equal access to economic resources to a broad cross-section of society. Although this definition is far from requiring equality of opportunity in society, it implies that societies where only a very small fraction of the population have well-enforced property rights do not have good economic institutions. Consequently, as we will see in some of the historical cases discussed below, a given set of economic institutions may have very different implications for economic growth depending on the technological possibilities and opportunities.

Given this definition of good economic institutions as providing secure property rights for a broad cross-section of society, our framework leads to a number of important com-

parative statics, and thus to an answer to our basic question. First, political institutions that place checks on those who hold political power, for example, by creating a balance of power in society, are useful for the emergence of good economic institutions. This result is intuitive; without checks on political power, power holders are more likely to opt for a set of economic institutions that are beneficial for themselves and detrimental for the rest of society, which will typically fail to protect property rights of a broad cross-section of people. Second, good economic institutions are more likely to arise when political power is in the hands of a relatively broad group with significant investment opportunities. The reason for this result is that, everything else equal, in this case power holders will themselves benefit from secure property rights.[2] Third, good economic institutions are more likely to arise and persist when there are only limited rents that power holders can extract from the rest of society, since such rents would encourage them to opt for a set of economic institutions that make the expropriation of others possible. These comparative statics therefore place political institutions at the center of the story, as emphasized by our term "hierarchy of institutions" above. Political institutions are essential both because they determine the constraints on the use of (de facto and de jure) political power and also which groups hold de jure political power in society. We will see below how these comparative statics help us understand institutional differences across countries and over time in a number of important historical examples.

1.3 Outline

In the next section we discuss how economic institutions constitute the basis for a fundamental theory of growth, and we contrast this with other potential fundamental theories. In section 3 we consider some empirical evidence that suggests a key role for economic institutions in determining long-run growth. We also emphasize some of the key problems involved in establishing a causal relationship between economic institutions and growth. We then show in section 4 how the experience of European colonialism can be used as a 'natural experiment' which can address these problems. Having established the central causal role of economic institutions and their importance relative to other factors in cross-country differences in economic performance, the rest of the paper focuses on developing a theory of economic institutions. Section 5 discusses four types of explanation for why countries have different institutions, and argues that

[2]The reason why we inserted the caveat of "a relatively broad group" is that when a small group with significant investment opportunities holds power, they may sometimes opt for an oligarchic system where their own property rights are protected, but those of others are not (see Acemoglu, 2003b).

the most plausible is the *social conflict view*. According to this theory, bad institutions arise because the groups with political power benefit from bad institutions. The emphasis on social conflict arises naturally from our observation above that economic institutions influence the distribution of resources as well as efficiency. Different groups or individuals will therefore prefer different institutions and conflict will arise as each tries to get their own way. Section 6 delves deeper into questions of efficiency and asks why a political version of the Coase Theorem does not hold. We emphasize the idea that commitment problems are intrinsic to the exercise of political power. In section 7 we argue that a series of historical examples of diverging economic institutions are best explained by the social conflict view. These examples illustrate how economic institutions are determined by the distribution of political power, and how this distribution is influenced by political institutions. Section 8 puts these ideas together to build our theory of institutions. In section 9 we then consider two more extended examples of the theory in action, the rise of constitutional rule in early modern Europe, and the creation of mass democracy, particularly in Britain, in the nineteenth and twentieth centuries. Section 10 concludes with a discussion of where this research program can go next.

2 Fundamental Causes of Income Differences

We begin by taking a step back. The presumption in the introduction was that economic institutions matter, and should in fact be thought of as one of the key fundamental causes of economic growth and cross-country differences in economic performance. How do we know this?

2.1 Three Fundamental Causes

If standard economic models of factor accumulation and endogenous technical change only provide proximate explanations of comparative growth, what types of explanations would constitute fundamental ones? Though there is no conventional wisdom on this, we can distinguish three such theories: the first set of theories, our main focus in this chapter, emphasize the importance of economic institutions, which influence economic outcomes by shaping economic incentives; the second emphasize geography, and the third emphasize the importance of culture (a fourth possibility is that differences are due to "luck," some societies were just lucky; however we do not believe that differences in luck by themselves constitute a sufficient fundamental causes of cross-country income differences).

2.1.1 Economic Institutions

At its core, the hypothesis that differences in economic institutions are the fundamental cause of different patterns of economic growth is based on the notion that it is the way that humans themselves decide to organize their societies that determines whether or not they prosper. Some ways of organizing societies encourage people to innovate, to take risks, to save for the future, to find better ways of doing things, to learn and educate themselves, solve problems of collective action and provide public goods. Others do not.

The idea that the prosperity of a society depends on its economic institutions goes back at least to Adam Smith, for example in his discussions of mercantilism and the role of markets, and was prominent in the work of many nineteenth century scholars such as John Stuart Mill (see the discussion in Jones, 1981): societies are economically successful when they have 'good' economic institutions and it is these institutions that are the *cause* of prosperity. We can think of these good economic institutions as consisting of an inter-related cluster of things. There must be enforcement of property rights for a broad cross-section of society so that all individuals have an incentive to invest, innovate and take part in economic activity. There must also be some degree of equality of opportunity in society, including such things as equality before the law, so that those with good investment opportunities can take advantage of them.[3]

One could think of other types of economic institutions, for instance markets. Traditional accounts of economic growth by historians, following the lead of Adam Smith, emphasized the spread of markets (Pirenne, 1937, Hicks, 1969) and more recent theories of comparative development are also based on differences in various economic institutions. Models of poverty traps in the tradition of Rosenstein-Rodan (1943), Murphy, Vishny and Shleifer (1989a,b) and Acemoglu (1995, 1997), are based on the idea that market imperfections can lead to the existence of multiple Pareto-ranked equilibria. As a consequence a country can get stuck in a Pareto inferior equilibrium, associated with poverty, but getting out of such a trap necessitates coordinated activities that the market cannot deliver. The literature initiated by Banerjee and Newman (1993) and Galor and Zeira (1993) is based on the idea that when capital markets are imperfect, the distribution of wealth matters for who can invest and societies with skewed income distributions can be stuck in poverty.

[3]In Acemoglu, Johnson and Robinson (2001), we coined the term *institutions of private property* for a cluster of would economic institutions, including the rule of law and the enforcement of property rights, and the term *extractive institutions* to designate institutions under which the rule of law and property rights are absent for large majorities of the population.

These theories provide interesting models of how incentives depend on expectations of others' behavior or the distribution of wealth given an underlying set of market imperfections. They take the market structure largely as given, however. We believe that the structure of markets is endogenous, and partly determined by property rights. Once individuals have secure property rights and there is equality of opportunity, the incentives will exist to create and improve markets (even though achieving perfect markets would be typically impossible). Thus we expect differences in markets to be an outcome of differing systems of property rights and political institutions, not unalterable characteristics responsible for cross-country differences in economic performance. This motivates our focus on economic institutions related to the enforcement of the property rights of a broad cross-section of society.

2.1.2 Geography

While institutional theories emphasize the importance of man-made factors shaping incentives, an alternative is to focus on the role of "nature", that is, on the physical and geographical environment. In the context of understanding cross-country differences in economic performance, this approach emphasizes differences in geography, climate and ecology that determine both the preferences and the opportunity set of individual economic agents in different societies. We refer to this broad approach as the "geography hypothesis". There are at least three main versions of the geography hypothesis, each emphasizing a different mechanism for how geography affects prosperity.

First, climate may be an important determinant of work effort, incentives, or even productivity. This idea dates back at least to the famous French philosopher, Montesquieu ([1748], 1989), who wrote in his classic book The Spirit of the Laws: "The heat of the climate can be so excessive that the body there will be absolutely without strength. So, prostration will pass even to the spirit; no curiosity, no noble enterprise, no generous sentiment; inclinations will all be passive there; laziness there will be happiness," and "People are ... more vigorous in cold climates. The inhabitants of warm countries are, like old men, timorous; the people in cold countries are, like young men, brave". One of the founders of modern economics Marshall is another prominent figure who emphasized the importance of climate, arguing: "vigor depends partly on race qualities: but these, so far as they can be explained at all, seem to be chiefly due to climate" (1890, p. 195).

Second, geography may determine the technology available to a society, especially in agriculture. This view is developed by an early Nobel Prize winner in economics,

Myrdal, who wrote "serious study of the problems of underdevelopment ... should take into account the climate and its impacts on soil, vegetation, animals, humans and physical assets—in short, on living conditions in economic development" (1968, volume 3, p. 2121). More recently, Diamond espouses this view, "... proximate factors behind Europe's conquest of the Americas were the differences in all aspects of technology. These differences stemmed ultimately from Eurasia's much longer history of densely populated ... [societies dependent on food production]," which was in turn determined by geographical differences between Europe and the Americas (1997, p. 358). The economist Sachs has been a recent and forceful proponent of the importance of geography in agricultural productivity, stating that "By the start of the era of modern economic growth, if not much earlier, temperate-zone technologies were more productive than tropical-zone technologies ..." (2001, p. 2).

The third variant of the geography hypothesis, especially popular over the past decade, links poverty in many areas of the world to their "disease burden," emphasizing that: "The burden of infectious disease is similarly higher in the tropics than in the temperate zones" (Sachs, 2000, p. 32). Bloom and Sachs (1998) claim that the prevalence of malaria, a disease which kills millions of children every year in sub-Saharan Africa, reduces the annual growth rate of sub-Saharan African economies by more than 1.3 percent a year (this is a large effect, implying that had malaria been eradicated in 1950, income per capita in sub-Saharan Africa would be double of what it is today).

2.1.3 Culture

The final fundamental explanation for economic growth emphasizes the idea that different societies (or perhaps different races or ethnic groups) have different cultures, because of different shared experiences or different religions. Culture is viewed as a key determinant of the values, preferences and beliefs of individuals and societies and, the argument goes, these differences play a key role in shaping economic performance.

At some level, culture can be thought to influence equilibrium outcomes for a given set of institutions. Possibly there are multiple equilibria connected with any set of institutions and differences in culture mean that different societies will coordinate on different equilibria. Alternatively, as argued by Greif (1993), different cultures generate different sets of beliefs about how people behave and this can alter the set of equilibria for a given specification of institutions (for example, some beliefs will allow punishment strategies to be used whereas others will not).

The most famous link between culture and economic development is that proposed by Weber (1930) who argued that the origins of industrialization in western Europe could be traced to the Protestant reformation and particularly the rise of Calvinism. In his view, the set of beliefs about the world that was intrinsic to Protestantism were crucial to the development of capitalism. Protestantism emphasized the idea of predestination in the sense that some individuals were 'chosen' while others were not. "We know that a part of humanity is saved, the rest damned. To assume that human merit or guilt play a part in determining this destiny would be to think of God's absolutely free decrees, which have been settled from eternity, as subject to change by human influence, an impossible contradiction" (Weber, 1930, p. 60).

But who had been chosen and who not? Calvin did not explain this. Weber (1930, p. 66) notes "Quite naturally this attitude was impossible for his followers ... for the broad mass of ordinary men ... So wherever the doctrine of predestination was held, the question could not be suppressed whether there was any infallible criteria by which membership of the *electi* could be known." Practical solutions to this problem were quickly developed, "... in order to attain that self-confidence intense worldly activity is recommended as the most suitable means. It and it alone disperses religious doubts and gives the certainly of grace" Weber (1930, pp. 66-67).

Thus "however useless good works might be as a means of attaining salvation ... nevertheless, they are indispensable as a sign of election. They are the technical means, not of purchasing salvation, but of getting rid of the fear of damnation" (p. 69). Though economic activity was encouraged, enjoying the fruits of such activity was not. "Waste of time is ... the first and in principle the deadliest of sins. The span of human life is infinitely short and precious to make sure of one's own election. Loss of time through sociability, idle talk, luxury, even more sleep than is necessary for health ... is worthy of absolute moral condemnation ... Unwillingness to work is symptomatic of the lack of grace" (pp. 104-105).

Thus Protestantism led to a set of beliefs which emphasized hard work, thrift, saving, and where economic success was interpreted as consistent with (if not actually signalling) being chosen by God. Weber contrasted these characteristics of Protestantism with those of other religions, such as Catholicism, which he argued did not promote capitalism. For instance on his book on Indian religion he argued that the caste system blocked capitalist development (Weber, 1958, p. 112).

More recently, scholars, such as Landes (1998), have also argued that the origins of Western economic dominance are due to a particular set of beliefs about the world and

how it could be transformed by human endeavor, which is again linked to religious differences. Although Barro and McCleary (2003) provide evidence of a positive correlation between the prevalence of religious beliefs, notably about hell and heaven, and economic growth, this evidence does not show a causal effect of religion on economic growth, since religious beliefs are endogenous both to economic outcomes and to other fundamental causes of income differences (points made by Tawney, 1926, and Hill, 1961b, in the context of Weber's thesis).

Ideas about how culture may influence growth are not restricted to the role of religion. Within the literature trying to explain comparative development there have been arguments that there is something special about particular cultural endowments, usually linked to particular nation states. For instance, Latin America may be poor because of its Iberian heritage, while North America is prosperous because of its Anglo-Saxon heritage (Véliz, 1994). In addition, a large literature in anthropology argues that societies may become 'dysfunctional' or 'maladapted' in the sense that they adopt a system of beliefs or ways or operating which do not promote the success or prosperity of the society (see Edgerton, 1992, for a survey of this literature). The most famous version of such an argument is due to Banfield (1958) who argued that the poverty of Southern Italy was due to the fact that people had adopted a culture of "amoral familiarism" where they only trusted individuals of their own families and refused to cooperate or trust anyone else. This argument was revived in the extensive empirical study of Putnam (1993) who characterized such societies as lacking "social capital". Although Putnam and others, for example, Knack and Keefer (1997) and Durlauf and Fafchamps (2003), document positive correlations between measures of social capital and various economic outcomes, there is no evidence of a causal effect, since, as with religious beliefs discussed above, measures of social capital are potentially endogenous.

3 Institutions Matter

We now argue that there is convincing empirical support for the hypothesis that differences in economic institutions, rather than geography or culture, *cause* differences in incomes per-capita. Consider first Figure 1.

This shows the cross-country bivariate relationship between the log of GDP per-capita in 1995 and a broad measure of property rights, "protection against expropriation risk", averaged over the period 1985 to 1995. The data on economic institutions come from Political Risk Services, a private company which assesses the risk that investments will be expropriated in different countries. These data, first used by Knack and Keefer

(1995) and subsequently by Hall and Jones (1999) and Acemoglu, Johnson and Robinson (2001, 2002a) are imperfect as a measure of economic institutions, but the findings are robust to using other available measures of economic institutions. The scatter plot shows that countries with more secure property rights, i.e., better economic institutions, have higher average incomes.

It is tempting to interpret Figure 1 as depicting a causal relationship (i.e., as establishing that secure property rights cause prosperity). Nevertheless, there are well known problems with making such an inference. First, there could be reverse causation – perhaps only countries that are sufficiently wealthy can afford to enforce property rights. More importantly, there might be a problem of omitted variable bias. It could be something else, e.g., geography, that explains both why countries are poor and why they have insecure property rights. Thus if omitted factors determine institutions and incomes, we would spuriously infer the existence of a causal relationship between economic institutions and incomes when in fact no such relationship exists. Trying to estimate the relationship between institutions and prosperity using Ordinary Least Squares, as was done by Knack and Keefer (1995) and Barro (1997) could therefore result in biased regression coefficients.

To further illustrate these potential *identification* problems, suppose that climate, or geography more generally, matters for economic performance. In fact, a simple scatterplot shows a positive association between latitude (the absolute value of distance from the equator) and income per capita. Montesquieu, however, not only claimed that warm climate makes people lazy and thus unproductive, but also unfit to be governed by democracy. He argued that despotism would be the political system in warm climates. Therefore, a potential explanation for the patterns we see in Figure 1 is that there is an omitted factor, geography, which explains both economic institutions and economic performance. Ignoring this potential third factor would lead to mistaken conclusions.

Even if Montesquieu's story appears both unrealistic and condescending to our modern sensibilities, the general point should be taken seriously: the relationship shown in Figure 1, and for that matter that shown in Figure 2, is not causal. As we pointed out in the context of the effect of religion or social capital on economic performance, these types of scatterplots, correlations, or their multidimensional version in OLS regressions, *cannot* establish causality.

What can we do? The solution to these problems of inference is familiar in microeconometrics: find a source of variation in economic institutions that should have no effect on economic outcomes, or depending on the context, look for a natural experiment.

As an example, consider first one of the clearest natural experiments for institutions.

3.1 The Korean Experiment

Until the end of World War II, Korea was under Japanese occupation. Korean independence came shortly after the Japanese Emperor Hirohito announced the Japanese surrender on August 15, 1945. After this date, Soviet forces entered Manchuria and North Korea and took over the control of these provinces from the Japanese. The major fear of the United States during this time period was the takeover of the entire Korean peninsular either by the Soviet Union or by communist forces under the control of the former guerrilla fighter, Kim Il Sung. U.S. authorities therefore supported the influential nationalist leader Syngman Rhee, who was in favor of separation rather than a united communist Korea. Elections in the South were held in May 1948, amidst a widespread boycott by Koreans opposed to separation. The newly elected representatives proceeded to draft a new constitution and established the Republic of Korea to the south of the 38th parallel. The North became the Democratic People's Republic of Korea, under the control of Kim Il Sung. These two independent countries organized themselves in very different ways and adopted completely different sets of institutions. The North followed the model of Soviet socialism and the Chinese Revolution in abolishing private property of land and capital. Economic decisions were not mediated by the market, but by the communist state. The South instead maintained a system of private property and the government, especially after the rise to power of Park Chung Hee in 1961, attempted to use markets and private incentives in order to develop the economy.

Before this "natural experiment" in institutional change, North and South Korea shared the same history and cultural roots. In fact, Korea exhibited an unparalleled degree of ethnic, linguistic, cultural, geographic and economic homogeneity. There are few geographic distinctions between the North and South, and both share the same disease environment. For example, the CIA Factbook describes the climate of North Korea as "temperate with rainfall concentrated in summer" and that of South Korea as "temperate, with rainfall heavier in summer than winter". In terms of terrain North Korea is characterized as consisting of "mostly hills and mountains separated by deep, narrow valleys; coastal plains wide in west, discontinuous in east," while South Korea is "mostly hills and mountains; wide coastal plains in west and south". In terms of natural resources North Korea is better endowed with significant reserves of coal, lead, tungsten, zinc, graphite, magnesite, iron ore, copper, gold, pyrites, salt, fluorspar, hydropower. South Korea's natural resources are "coal, tungsten, graphite, molybdenum,

lead, hydropower potential." Both countries share the same geographic possibilities in terms of access to markets and the cost of transportation.

Other man-made initial economic conditions were also similar, and if anything, advantaged the North. For example, there was significant industrialization during the colonial period with the expansion of both Japanese and indigenous firms. Yet this development was concentrated more in the North than the South. For instance, the large Japanese zaibatsu of Noguchi, which accounted for one third of Japanese investment in Korea, was centered in the North. It built large hydroelectric plants, including the Suiho dam on the Yalu river, second in the world only to the Boulder dam on the Colorado river. It also created Nippon Chisso, the second largest chemical complex in the world that was taken over by the North Korean state. Finally, in Ch'ongjin North Korea also had the largest port on the Sea of Japan. All in all, despite some potential advantages for the North,[4] Maddison (2001) estimates that at the time of separation, North and South Korea had approximately the same income per capita.

We can therefore think of the splitting on the Koreas 50 years ago as a natural experiment that we can use to identify the causal influence of a particular dimension of institutions on prosperity. Korea was split into two, with the two halves organized in radically different ways, and with geography, culture and many other potential determinants of economic prosperity held fixed. Thus any differences in economic performance can plausibly be attributed to differences in institutions.

Consistent with the hypothesis that it is institutional differences that drive comparative development, since separation, the two Koreas have experienced dramatically diverging paths of economic development: see Figure 3. By the late 1960's South Korea was transformed into one of the Asian "miracle" economies, experiencing one of the most rapid surges of economic prosperity in history while North Korea stagnated. By 2000 the level of income in South Korea was $16,100 while in North Korea it was only $1,000. By 2000 the South had become a member of the Organization of Economic Cooperation and Development, the rich nations club, while the North had a level of per-capita income about the same as a typical sub-Saharan African country. There is only one plausible explanation for the radically different economic experiences on the two Koreas after 1950: their very different institutions led to divergent economic outcomes. In this context, it is noteworthy that the two Koreas not only shared the same geography, but also the same culture.

[4]Such initial differences were probably eradicated by the intensive bombing campaign that the United States unleashed in the early 1950's on North Korea (see Cumings, 2004, chapter 1).

It is possible that Kim Il Sung and Communist Party members in the North believed that communist policies would be better for the country and the economy in the late 1940s. However, by the 1980s it was clear that the communist economic policies in the North were not working. The continued efforts of the leadership to cling to these policies and to power can only be explained by those leaders wishing to look after their own interests at the expense of the population at large. Bad institutions are therefore kept in place, clearly not for the benefit of society as a whole, but for the benefit of the ruling elite, and this is a pattern we encounter in most cases of institutional failure that we discuss in detail below.

However convincing on its own terms, the evidence from this natural experiment is not sufficient for the purposes of establishing the importance of economic institutions as the primary factor shaping cross-country differences in economic prosperity. First, this is only one case, and in the better-controlled experiments in the natural sciences, a relatively large sample is essential. Second, here we have an example of an extreme case, the difference between a market-oriented economy and a communist one. Few social scientists today would deny that a lengthy period of totalitarian centrally planned rule has significant economic costs. And yet, many might argue that differences in economic institutions among capitalist economies or among democracies are not the major factor leading to differences in their economic trajectories. To establish the major role of economic institutions in the prosperity and poverty of nations we need to look at a larger scale "natural experiment" in institutional divergence.

3.2 The Colonial Experiment

The colonization of much of the world by Europeans provides such a large scale natural experiment. Beginning in the early fifteenth century and massively intensifying after 1492, Europeans conquered many other nations. The colonization experience transformed the institutions in many diverse lands conquered or controlled by Europeans. Most importantly, Europeans imposed very different sets of institutions in different parts of their global empire, as exemplified most sharply by the contrast to the economic institutions in the northeast of America to those in the plantation societies of the Caribbean. As a result, while geography was held constant, Europeans initiated large changes in economic institutions, in the social organization of different societies. We will now show that this experience provides evidence which conclusively establishes the central role of economic institutions in development. Given the importance of this material and the details we need to provide, we discuss the colonial experience in the

next section.

4 The Reversal of Fortune

The impact of European colonialism on economic institutions is perhaps most dramatically conveyed by a single fact—historical evidence shows that there has been a remarkable Reversal of Fortune in economic prosperity within former European colonies. Societies like the Mughals in India, and the Aztecs and the Incas in the Americas were among the richest civilizations in 1500, yet the nation states that now coincide with the boundaries of these empires are among the poorer societies of today. In contrast, countries occupying the territories of the less-developed civilizations in North America, New Zealand and Australia are now much richer than those in the lands of the Mughals, Aztecs and Incas.

4.1 The reversal among the former colonies

The Reversal of Fortune is not confined to such comparisons. Using reasonable proxies for prosperity before modern times, we can show that it is a much more systematic phenomenon. Our proxies for income per capita in pre-industrial societies are urbanization rates and population density. Only societies with a certain level of productivity in agriculture and a relatively developed system of transport and commerce can sustain large urban centers and a dense population. Figure 4 shows the relationship between income per capita and urbanization (fraction of the population living in urban centers with greater than 5,000 inhabitants) today, and demonstrates that in the current period there is a significant relationship between urbanization and prosperity.

Naturally, high rates of urbanization do not mean that the majority of the population lived in prosperity. In fact, before the twentieth century urban centers were often centers of poverty and ill health. Nevertheless, urbanization is a good proxy for average income per capita in society, which closely corresponds to the measure we are using to look at prosperity.

Figures 5 and 6 show the relationship between income per capita today and urbanization rates and (log) population density in 1500 for the sample of European colonies.[5] We pick 1500 since it is before European colonization had an effect on any of these

[5]The sample includes the countries colonized by the Europeans between the 15th and the 19th centuries as part of their overseas expansion after the discovery of the New World and the rounding of the Cape of Good Hope. It therefore excludes Ireland, parts of the Russian Empire and also the Middle East and countries briefly controlled by European powers as U.N. Mondays during the 20th century.

societies. A strong negative relationship, indicating a reversal in the rankings in terms of economic prosperity between 1500 and today, is clear in both figures. In fact, the figures show that in 1500 the temperate areas were generally less prosperous than the tropical areas, but this pattern too was reversed by the twentieth century.

The urbanization data for these Figures come from Bairoch (1988), Bairoch, Batou and Chèvre (1988), Chandler (1987), and Eggimann (1999). The data on population density are from McEvedy and Jones (1978). Details and further results are in Acemoglu, Johnson and Robinson (2002a).

There is something extraordinary about this reversal. For example, after the initial spread of agriculture there was remarkable persistence in urbanization and population density for all countries, including those which were to be subsequently colonized by Europeans. In Figures 7 and 8 we show the relationships for urbanization plotting separately the relationship between urbanization in 1000 and in 1500 for the samples of colonies and all other countries. Both figures show persistence, not reversal. Although Ancient Egypt, Athens, Rome, Carthage and other empires rose and fell, what these pictures show is that there was remarkable persistence in the prosperity of regions.

Moreover, reversal was not the general pattern in the world after 1500. Figure 9 shows that within countries not colonized by Europeans in the early modern and modern period, there was no reversal between 1500 and 1995. There is therefore no reason to think that what is going on in Figures 5 and 6 is some sort of natural reversion to the mean.

4.2 Timing of the Reversal

When did the reversal occur? One possibility is that it arose shortly after the conquest of societies by Europeans but Figures 10 and 11 show that the previously-poor colonies surpassed the former highly-urbanized colonies starting in the late eighteenth and early nineteenth centuries, and this went hand in hand with industrialization. Figure 10 shows average urbanization in colonies with relatively low and high urbanization in 1500. The initially high-urbanization countries have higher levels of urbanization and prosperity until around 1800. At that time the initially low-urbanization countries start to grow much more rapidly and a prolonged period of divergence begins. Figure 11 shows industrial production per capita in a number of countries. Although not easy to see in the figure, there was more industry (per capita and total) in India in 1750 than in the United States. By 1860, the United States and British colonies with relatively good economic institutions, such as Australia and New Zealand, began to move ahead

rapidly, and by 1953, a huge gap had opened up.

4.3 Interpreting the Reversal

Which of the three broad hypotheses about the sources of cross-country income differences are consistent with the reversal and its timing? These patterns are clearly inconsistent with simple geography based views of relative prosperity. In 1500 it was the countries in the tropics which were relatively prosperous, in 2003 it is the reverse. This makes it implausible to base a theory of relative prosperity today, as Sachs (2000, 2001) does, on the intrinsic poverty of the tropics. This argument is inconsistent with the historical evidence.

Nevertheless, following Diamond (1997), one could propose what Acemoglu, Johnson and Robinson (2002a) call a "sophisticated geography hypothesis" which claims that geography matters but in a time varying way. For example, Europeans created "latitude specific" technology, such as heavy metal ploughs, that only worked in temperate latitudes and not with tropical soils. Thus when Europe conquered most of the world after 1492, they introduced specific technologies that functioned in some places (the United States, Argentina, Australia) but not others (Peru, Mexico, West Africa). However, the timing of the reversal, coming as it does in the nineteenth century, is inconsistent with the most natural types of sophisticated geography hypotheses. Europeans may have had latitude specific technologies, but the timing implies that these technologies must have been industrial, not agricultural, and it is difficult to see why industrial technologies do not function in the tropics (and in fact, they have functioned quite successfully in tropical Singapore and Hong Kong).[6]

Similar considerations weigh against the culture hypothesis. Although culture is slow-changing the colonial experiment was sufficiently radical to have caused major changes in the cultures of many countries that fell under European rule. In addition, the destruction of many indigenous populations and immigration from Europe are likely to have created new cultures or at least modified existing cultures in major ways (see Vargas Llosa, 1989, for a fictionalized account of just such a cultural change). Nevertheless, the culture hypothesis does not provide a natural explanation for the reversal, and has nothing to say on the timing of the reversal. Moreover, we discuss below how econometric models that control for the effect of institutions on income do not find any evidence of an effect

[6]A possible link is that proposed by Lewis (1978) who argued that tropical agriculture is less productive than temperate agriculture, and that an 'agricultural revolution' is a prerequisite to an industrial revolution because high agricultural productivity is needed to stimulate the demand for industrial goods.

of religion or culture on prosperity.

The most natural explanation for the reversal comes from the institutions hypothesis, which we discuss next.

4.4 Economic Institutions and the Reversal

Is the Reversal of Fortune consistent with a dominant role for economic institutions in comparative development? The answer is yes. In fact, once we recognize the variation in economic institutions created by colonization, we see that the Reversal of Fortune is exactly what the institutions hypothesis predicts.

In Acemoglu, Johnson and Robinson (2002a) we tested the connection between initial population density, urbanization, and the creation of good economic institutions. We showed that, others things equal, the higher the initial population density or the greater initial urbanization, the worse were subsequent institutions, including both institutions right after independence and today. Figures 12 and 13 show these relationships using the same measure of current economic institutions used in Figure 1, protection against expropriation risk today. They document that the relatively densely settled and highly urbanized colonies ended up with worse (or 'extractive') institutions, while sparsely-settled and non-urbanized areas received an influx of European migrants and developed institutions protecting the property rights of a broad cross-section of society. European colonialism therefore led to an institutional reversal, in the sense that the previously-richer and more-densely settled places ended up with worse institutions.[7]

To be fair, it is possible that the Europeans did not actively introduce institutions discouraging economic progress in many of these places, but inherited them from previous civilizations there. The structure of the Mughal, Aztec and Inca empires were already very hierarchical with power concentrated in the hands of narrowly based ruling elites and structured to extract resources from the majority for the benefit of a minority. Often Europeans simply took over these existing institutions. Whether this is so is secondary for our focus, however. What matters is that in densely-settled and relatively-developed places it was in the interests of Europeans to have institutions facilitating the extraction of resources thus not respecting the property rights of the majority, while in the sparsely-settled areas it was in their interests to develop institutions protecting

[7]The institutional reversal does not mean that institutions were necessarily better in the previously more densely-settled areas (see the next paragraph). It only implies a tendency for the relatively poorer and less densely-settled areas to end up with better institutions than previously-rich and more densely-settled areas.

property rights. These incentives led to an institutional reversal.

The institutional reversal, combined with the institutions hypothesis, predicts the Reversal of Fortune: relatively rich places got relatively worse economic institutions, and if these institutions are important, we should see them become relatively poor over time. This is exactly what we find with the Reversal of Fortune.

Moreover, the institutions hypothesis is consistent with the timing of the reversal. Recall that the institutions hypothesis links incentives to invest in physical and human capital and in technology to economic institutions, and argues that economic prosperity results from these investments. Therefore, economic institutions should become more important when there are major new investment opportunities. The opportunity to industrialize was the major investment opportunity of the nineteenth century. Countries that are rich today, both among the former European colonies and other countries, are those that industrialized successfully during this critical period.

4.5 Understanding the colonial experience

The explanation for the reversal that emerges from our discussion so far is one in which the economic institutions in various colonies were shaped by Europeans to benefit themselves. Moreover, because conditions and endowments differed between colonies, Europeans consciously created different economic institutions, which persisted and continue to shape economic performance. Why did Europeans introduce better institutions in previously-poor and unsettled areas than in previously-rich and densely-settled areas? The answer to this question relates to the comparative statics of our theoretical framework. Leaving a full discussion to later, we can note a couple of obvious ideas.

Europeans were more likely to introduce or maintain economic institutions facilitating the extraction of resources in areas where they would benefit from the extraction of resources. This typically meant areas controlled by a small group of Europeans, and areas offering resources to be extracted. These resources included gold and silver, valuable agricultural commodities such as sugar, but most importantly people. In places with a large indigenous population, Europeans could exploit the population, be it in the form of taxes, tributes or employment as forced labor in mines or plantations. This type of colonization was incompatible with institutions providing economic or civil rights to the majority of the population. Consequently, a more developed civilization and a denser population structure made it more profitable for the Europeans to introduce worse economic institutions.

In contrast, in places with little to extract, and in sparsely-settled places where the

Europeans themselves became the majority of the population, it was in their interests to introduce economic institutions protecting their own property rights.

4.6 Settlements, Mortality and Development

The initial conditions we have emphasized so far refer to indigenous population density and urbanization. In addition, the disease environments differed markedly among the colonies, with obvious consequences on the attractiveness of European settlement. As we noted above, when Europeans settled, they established institutions that they themselves had to live under. Therefore, whether Europeans could settle or not had an exogenous effect on the subsequent path of institutional development. In other words, if the disease environment 200 or more years ago affects outcomes today only through its effect on institutions today, then we can use this historical disease environment as an exogenous source of variation in current institutions. From an econometric point of view we have a valid instrument which will enable us to pin down the casual effect of economic institutions on prosperity.[8]

We developed this argument in Acemoglu, Johnson and Robinson (2001) and investigated it empirically. We used initial conditions in the European colonies, particularly data from Curtin (1989, 1998) and Gutierrez (1986) on the mortality rates faced by Europeans (primarily soldiers, sailors, and bishops), as instruments for current economic institutions. The justification for this is that, outside of its effect on economic institutions during the colonial period, historical European mortality has no impact on current income levels. Figures 14 and 15 give scatter plots of this data against contemporaneous economic institutions and GDP per-capita. The sample is countries which were colonized by Europeans in the early modern and modern periods and thus excludes, among others, China, Japan, Korea, Thailand.

Figure 14 shows the very strong relationship between the historical mortality risk faced by Europeans and the current extent to which property rights are enforced. A bivariate regression has an R^2 of 0.26. It also shows that there were very large differences in European mortality. Countries such as Australia, New Zealand and the United States were very healthy with life expectancy typically greater than in Britain. On the other hand mortality was extremely high in Africa, India and South-East Asia. Differential

[8]Although European mortality is potentially correlated with indigenous mortality, which may determine income today, in practice local populations have developed much greater immunity to malaria and yellow fever. Thus the historical experience of European mortality is a valid instrument for institutional development. See Acemoglu, Johnson and Robinson (2001).

mortality was largely due to tropical diseases such as malaria and yellow fever and at the time it was not understood how these diseases arose nor how they could be prevented or cured.

In Acemoglu, Johnson and Robinson (2001) we showed, using European mortality as an instrument for the current enforcement of property rights, that most of the gap between rich and poor countries today is due to differences in economic institutions. More precisely, we showed (p. 1387) that if one took two typical—in the sense that they both lie on the regression line—countries with high and low expropriation risk, like Nigeria and Chile, then almost the entire difference in incomes per-capita between them could be explained by the differences in the security of property rights. We also presented regression evidence that showed that once the effect of economic institutions on GDP per-capita was properly controlled for, geographical variables, such as latitude, whether or not a country is land-locked and the current disease environment, have no explanatory power for current prosperity.

These ideas and results provide an interpretation of why there are strong correlations between geographical variables such as latitude and income per-capita. Basically this is because Europeans did not have immunity to tropical diseases during the colonial period and thus settler colonies tended, other things equal, to be created in temperate latitudes. Thus the historical creation of economic institutions was correlated with latitude. Without considering the role of economic institutions it is easy to find a spurious relationship between latitude and income per-capita. However, once economic institutions are properly controlled for, these relationships go away. There is no causal effect of geography on prosperity today, though geography may have been important historically in shaping economic institutions.

What about the role of culture? On the face of it, the Reversal of Fortune is consistent with cultural explanations of comparative growth. The Europeans not only brought new institutions, they also brought their own cultures. There seem to be three main ways to test this idea. First, cultures may be systematically related to the national identity of the colonizing power. For example, the British may have implanted a 'good' Anglo-Saxon culture into colonies such as Australia and the United States, while the Spanish may have condemned Latin America by endowing it with a Hispanic or Iberian culture (the academic literature is full of ideas like this, for recent versions see Véliz, 1994, North, Summerhill and Weingast, 2000, and Wiarda, 2001). Second, following Landes (1998), Europeans may have had a culture, for example a work ethic or set of beliefs, which was uniquely propitious to prosperity. Finally, following Weber (1930),

Europeans also brought different religions with different implications for prosperity. Such a hypothesis could explain why Latin America is relatively poor since its citizens are primarily Roman Catholic, while North America is relatively rich because its citizens are mostly Protestant.

However, the econometric evidence in Acemoglu, Johnson and Robinson (2001) is not consistent with any these views. Once we control properly for the effects of economic institutions, neither the identity of the colonial power, nor the contemporary fraction of Europeans in the population, nor the proportions of the populations of various religions, are significant determinants of income per capita.

These econometric results are supported by historical examples. For instance, with respect to the identity of the colonizing power, in the 17th century the Dutch had perhaps the best domestic economic institutions in the world but the colonies they created in South-East Asia ended up with institutions designed for the extraction of resources, providing little economic or civil rights to the indigenous population.

It is also be clear that the British in no way simply re-created British institutions in their colonies. For example, by 1619 the North American colony of Virginia had a representative assembly with universal male suffrage, something that did not arrive in Britain itself until 1919. Another telling example is due to Newton (1914) and Kupperman (1993), who showed that the Puritan colony in Providence Island in the Caribbean quickly became just like any other Caribbean slave colony despite the Puritanical inheritance. Although no Spanish colony has been as successful economically as British colonies such as the United States, it is also important to note that Britain had many unsuccessful colonies (in terms of per capita income), such as in Africa, India and Bangladesh (see Acemoglu, Johnson and Robinson, 2004).

To emphasize that the culture or the religion of the colonizer was not at the root of the divergent economic performances of the colonies, Figure 16 shows the reversal among the British colonies (with population density in 1500 on the horizontal axis). Just as in Figure 6, there is a strong negative relationship between population density in 1500 and income per capita today.

With respect to the role of Europeans, Singapore and Hong Kong are now two of the richest countries in the world, despite having negligible numbers of Europeans. Moreover, Argentina and Uruguay have higher proportions of people of European descent than the United States and Canada, but are much less rich. To further document this, Figure 17 shows a similar reversal of fortune for countries where the fraction of those with European descent in 1975 is less than 5 percent of the population.

Overall, the evidence is not consistent with a major role of geography, religion or culture transmitted by the identity of the colonizer or the presence of Europeans. Instead, differences in economic institutions appear to be the robust causal factor underlying the differences in income per capita across countries. Institutions are therefore the fundamental cause of income differences and long-run growth.

5 Why Do Institutions Differ?

We saw that economic institutions matter, indeed are central in determining relative prosperity. In terms of the different fundamental theories that we discussed, there is overwhelming support for the emphasis of North and Thomas on institutions, as opposed to alternative candidate explanations which emphasize geography or culture. Yet, as we discussed in the introduction, finding that differences in economic institutions can account for the preponderance of differences in per-capita income between countries creates as many questions as it answers. For example, why do countries have different economic institutions? If poor countries are poor because they have bad economic institutions why do they not change them to better institutions? In short, to explain the evidence presented in the last two sections we need a theory of economic institutions. The theory will help to explain the equilibrium set of economic institutions in a particular country and the comparative statics of this theory will help to explain why economic institutions differ across countries.

In the Introduction (section 1.2), we began to develop such a theory based on social conflict over economic institutions. We have now substantiated the first point we made there, that economic institutions determine prosperity. We must now move to substantiate our second point, that economic institutions must be treated as endogenous and what which economic institutions emerge depends on the distribution of political power in society. This is a key step towards our theory of economic institutions. In the process of substantiating this point however it is useful to step back and discuss other alternative approaches to developing a theory of economic institutions. Broadly speaking, there are four main approaches to the question of why institutions differ across countries, one of which coincides with the approach we are proposing, the social conflict view. We next discuss each of these separately and our assessment as to whether they provide a satisfactory framework for thinking about differences in economic institutions (see Acemoglu, 2003a, and Robinson, 1998, for related surveys of some of these approaches). We shall conclude that the approach we sketched in section 1.2 is by far the most promising one.

5.1 The Efficient Institutions View—The Political Coase Theorem

According to this view, societies will choose the economic institutions that are socially efficient. How this surplus will be distributed among different groups or agents does not affect the choice of economic institutions. We stress here that the concept of efficiency is stronger than simply Pareto Optimality; it is associated with surplus, wealth or output maximization.

The underlying reasoning of this view comes from the Coase Theorem. Coase (1960) argued that when different economic parties could negotiate costlessly, they will be able to bargain to internalize potential externalities. A farmer, who suffers from the pollution created by a nearby factory, can pay the factory owner to reduce pollution. Similarly, if the current economic institutions benefit a certain group while creating a disproportionate cost for another, these two groups can negotiate to change the institutions. By doing so they will increase the size of the total surplus that they can divide between themselves, and they can then bargain over the distribution of this additional surplus.

Many different versions of the efficient economic institutions view have been proposed. Indeed, assuming that existing economic institutions are efficient is a standard methodological approach of economists, i.e., observing an institution, one tries to understand what are the circumstances that lead it to be efficient. Demsetz (1967) argued that private property emerged from common property when land became sufficiently scarce and valuable that it was efficient to privatize it. More recently, Williamson's (1985) research, as well as Coase's (1936) earlier work and the more formal analysis by Grossman and Hart (1986), argues that the governance of firms or markets is such as to guarantee efficiency (given the underlying informational and contractual constraints). Williamson argued that firms emerged as an efficient response to contractual problems that plague markets, particularly the fact that there may be ex-post opportunism when individuals make relationship specific investments. Another famous application of the efficient institutions view is due to North and Thomas (1973) who argued that feudal economic institutions, such as serfdom, were an efficient contract between serfs and lords. The lords provided a public good, protection, in exchange for the labor of the serfs on their lands. In this view, without a modern fiscal system this was an efficient way to organize this exchange. (See Townsend, 1993, for a recent version of the idea that other economic institutions of Medieval Europe, such as the open field system, were efficient).

Williamson and North and Thomas do not specify how different parties will reach agreement to achieve efficient economic institutions, and this may be problematical in

the sense that many economic institutions relevant for development are collective choices not individual bargains. There may therefore be free riding problems inherent in the creation of efficient economic institutions. Nevertheless, the underlying idea, articulated by Becker (1960) and Wittman (1989), is that, at least in democracies, competition among pressure groups and political parties will lead to efficient policies and collective choices. In their view, an inefficient economic institution cannot be stable because a political entrepreneur has an incentive to propose a better economic institution and with the extra surplus generated will be able to make himself more attractive to voters. The efficient institutions view regards the structure of political institutions or power as irrelevant. This may matter for the distribution of total surplus, but it will not matter for efficiency itself. The 'efficient' set of political institutions is therefore indeterminate.

The notion that a Coasian logic applies in political life as well as in economics is referred to by Acemoglu (2003a) as the Political Coase Theorem. Although the intuition that individuals and groups will strive towards efficient economic outcomes is appealing, there are both theoretical and empirical limits to the Political Coase Theorem. First, as argued by Acemoglu (2003a) and further discussed below, in politics there is an inherent commitment problem, often making the Political Coase Theorem inapplicable.

Second, the Political Coase Theorem does not take us very far in understanding the effect of economic (or indeed political) institutions on economic outcomes – in this view, economic institutions are chosen efficiently, and all societies have the best possible economic institutions given their needs and underlying structures; hence, with the Political Coase Theorem, economic institutions cannot be the fundamental cause of income differences. However, the empirical results we discussed above suggest a major role for such institutional differences.

The only way to understand these patterns is to think of economic institutions varying for reasons other than the underlying needs of societies. In fact, the instrumental variables and natural experiment strategies we exploited above make use precisely of a source of variation unrelated to the underlying needs of societies. For example, South and North Korea did not adopt very different economic systems because they had different needs, but because different systems were imposed on them for other exogenous reasons. In sum, we need a framework for understanding why certain societies consistently end up with economic institutions that are not, from a social point of view, in their best interests. We need a framework other than the Political Coase Theorem.

5.2 The Ideology View

A second view is that economic institutions differ across countries because of ideological differences – because of the similarity between this and the previous view, Acemoglu (2003a) calls this the Modified Political Coase Theorem. According to this view, societies may choose different economic institutions, with very different implications, because they—or their leaders—disagree about what would be good for the society. According to this approach, there is sufficient uncertainty about the right economic institutions that well-meaning political actors differ about what's good for their own people. Societies where the leaders or the electorate turn out to be right ex post are those that prosper. The important point is that, just as with the efficient institutions view, there are strong forces preventing the implementation of policies that are *known* to be bad for the society at large.

Several theoretical models have developed related ideas. For example, Piketty (1995) examined a model where different people have different beliefs about how much effort is rewarded in society. If effort is not rewarded then taxation generates few distortions and agents with such beliefs prefer a high tax rate. On the other hand if one believes that effort is rewarded then low taxes are preferable. Piketty showed that dispersion of beliefs could create dispersion of preferences over tax rates, even if all agents had the same objective. Moreover, incorrect beliefs could be self-fulfilling and persist over time because different beliefs tend to generate information consistent with those beliefs. Romer (2003) also presents a model where voters have different beliefs and showed that if mistakes are correlated, then society can choose a socially inefficient outcome. These models show that if different societies have different beliefs about what is socially efficient they can rationally choose different economic institutions.

Belief differences clearly do play a role in shaping policies and institutions. Several interesting examples of this come from the early experience of independence in former British colonies. For example, it is difficult to explain Julius Nyerere's policies in Tanzania without some reference to his and other leading politicians' beliefs about the desirability of a socialist society. It also appears true that in India the Fabian socialist beliefs of Jawaharlal Nehru were important in governing the initial direction that Indian economic policies took.

Nevertheless, the scope of a theory of institutional divergence and comparative development based on ideology seems highly limited. Can we interpret the differences in institutional development across the European colonies or the divergence in the eco-

nomic institutions and policies between the North and South of Korea as resulting from differences in beliefs? For example, could it be the case that while Rhee, Park, and other South Korean leaders believed in the superiority of capitalist institutions and private property rights enforcement, Kim Il Sung and Communist Party members in the North believed that communist policies would be better for the country?

In the case of South versus North Korea, this is certainly a possibility. However, even if differences in beliefs could explain the divergence in economic institutions in the immediate aftermath of separation, by the 1980s it was clear that the communist economic policies in the North were not working. The continued effort of the leadership to cling to these policies, and to power, can only be explained by leaders looking after their own interests at the expense of the population at large. Most likely, North Korean leaders, the Communist Party, and bureaucratic elites are prolonging the current system, which gives them greater economic and political returns than the alternative, even though they fully understand the costs that the system imposes on the North Korean people.

Differences in colonial policies are even harder to explain on the basis of differences in ideology. British colonists established different economic institutions in very different parts of the world: in the Caribbean they set up plantation societies based on slavery, supported by highly oppressive economic institutions. In contrast, the economic institutions that developed in areas where the British settled, and where there was no large population of indigenous to be captured and put to work, and where slavery could not be profitably used, such as northeastern United States, Canada, Australia and New Zealand, were very different. Moreover, differences in the incentives of the colonists in various colonies are easy to understand: when they did not settle, they were choosing economic institutions simply to extract resources from the native population. When they settled in large numbers, economic institutions and policies emerged in order to protect them in the future and encourage investment and prosperity.

These considerations make us tend towards a view which emphasizes the actions of key economic and political agents that are taken rationally and in recognition of their consequences, not simply differences in beliefs. We do not deny that belief differences and ideology often play important roles but we do not believe that a satisfactory theory of institutional differences can be founded on differences in ideology.

5.3 The Incidental Institutions View

The efficient institutions view is explicitly based on economic reasoning: the social costs and benefits of different economic institutions are weighed against each other to deter-

mine which economic institutions should prevail. Efficiency arises because individuals ultimately calculate according to social costs and benefits. Institutions are therefore choices. A different approach, popular among many political scientists and sociologists, but also some economists, is to downplay choices and to think of institutions, both economic and political, as the by-product or unintended consequence of other social interactions or historical accidents. In other words, historical accidents at critical junctures determine institutions, and these institutions persist for a long time, with significant consequences.

Here, we discuss two such theories. The first is the theory of political institutions developed by Moore (1966) in his *Social Origins of Dictatorship and Democracy*, the second is the recent emphasis in the economics literature on legal origins, for example as in the work of Shleifer and his co-authors (La Porta, Lopez-de-Silanes, Shleifer and Vishny, 1998, 1999, Djankov, LaPorta, Lopez-de-Silanes and Shleifer, 2002, 2003, Glaeser and Shleifer, 2002).

Moore attempted to explain the different paths of political development in Britain, Germany and Russia. In particular, he investigated why Britain evolved into a democracy, while Germany succumbed to fascism and Russia had a communist revolution. Moore stressed the extent of commercialization of agriculture and resulting labor relations in the countryside, the strength of the 'bourgeoisie,' and the nature of class coalitions. In his theory, democracy emerged when there was a strong, politically assertive, commercial middle class, and when agriculture had commercialized so that there were no feudal labor relations in the countryside. Fascism arose when the middle classes were weak and entered into a political coalition with landowners. Finally, a communist revolution resulted when the middle classes were non-existent, agriculture was not commercialized, and rural labor was repressed through feudal regulations. In Moore's theory, therefore, class coalitions and the way agriculture is organized determine which political institutions will emerge. However, the organization of agriculture is not chosen with an eye to its effects on political institutions, so these institutions are an unintended consequence. Although Moore was not explicitly concerned with economic development, it is a direct implication of his analysis that societies may end up with institutions that do not maximize income or growth, for example, when they take the path to communist revolution.

Beginning with the work on shareholder rights (La Porta, Lopez-de-Silanes, Shleifer and Vishny, 1998), continuing to the efficiency of government (La Porta, Lopez-de-Silanes, Shleifer and Vishny, 1999) and more recently the efficiency of the legal system

(Djankov, La Porta, Lopez-de-Silanes and Shleifer, 2003), Shleifer and his co-authors have argued that a central source of variation in many critical economic institutions is the origin of the legal system. For example, "Civil laws give investors weaker legal rights than common laws do, independent of the level of per-capita income. Common-law countries give both shareholders and creditors—relatively speaking—the strongest, and French-civil-law countries the weakest, protection. " (La Porta et al., 1998, p. 1116)

These differences have important implications for resource allocation. For example, when shareholders have poor protection of their rights, ownership of shares tends to be more highly concentrated. Djankov et al. (2003) collected a cross-national dataset on how different countries legal systems dealt with the issue of evicting a tenant for nonpayment of rent and collecting on a bounced check. They used these data to construct an index of procedural formalism of dispute resolution for each country and showed that such formalism was systematically greater in civil than in common law countries, and is associated with higher expected duration of judicial proceedings, less consistency, less honesty, less fairness in judicial decisions, and more corruption. Legal origins therefore seems to matter for important institutional outcomes.

Where do legal origins come from? The main argument is that they are historical accidents, mostly related to the incidence of European colonialism. For example, Latin American countries adopted the Napoleonic codes in the nineteenth century because these were more compatible with their Spanish legal heritage. Importantly, the fact that Latin American countries therefore have 'French legal origin' is due to a historical accident and can be treated as exogenous with respect to current institutional outcomes. What about the difference between common law and civil law? Glaeser and Shleifer (2002) argue that the divergence between these systems stems from the medieval period and reflects the balance of power between the lords and the king in England and France. Once these systems established, they persisted long after the initial rationale vanished.

Although we believe that historical accidents and persistence are important, in reality the aspect of choice over institutions seems too important to be denied. Even if institutions have a tendency to persist, their persistence is still a choice, in the sense that if the agents decided to change institutions, change would be possible. There are important examples from history of countries radically changing their legal systems such as in Japan after the Meiji restoration, Russia after the Crimean War, and Turkey under Mustafa Kemal in the 1920's. Another example might be central planning of the economy. Though many countries adopted this way or organizing the economy some abandoned it while others, such as North Korea and Cuba, still maintain it. The point

here is that though institutions may in some circumstances be the incidental outcome of history, at some point people will start to ask why society has the institutions that it does and to consider other alternatives. At this point we are back in the realm of choice.

5.4 The Social Conflict View

According to this view, economic (and political) institutions are not always chosen by the whole society (and not for the benefit of the whole society), but by the groups that control political power at the time (perhaps as a result of conflict with other groups). These groups will choose the economic institutions that maximize their own rents, and the economic institutions that result may not coincide with those that maximize total surplus, wealth or income. For example, economic institutions that enforce property rights by restricting state predation may not be in the interest of a ruler who wants to appropriate assets in the future. By establishing property rights, this ruler would be reducing his own future rents, so may well prefer economic institutions other than enforced private property. Therefore, equilibrium economic institutions will not be those that maximize the size of the overall pie, but the slice of the pie taken by the powerful groups.

The first systematic development of this point of view in the economics literature is North (1981), who argued in the chapter on "A Neoclassical Theory of the State" that agents who controlled the state should be modeled as self-interested. He then argued that the set of property rights that they would choose for society would be those that maximized their payoff and because of 'transactions costs,' these would not necessarily be the set that maximized social welfare. One problem with North's analysis is that he does not clarify what the transactions costs creating a divergence between the interests of the state and the citizens are. Here, we will argue that commitment problems are at the root of this divergence.

The notion that elites, i.e., the politically powerful, may opt for economic institutions which increase their incomes, often at the expense of society, is of course also present in much of the Marxist and dependency theory literature. For example, Dobb (1948), Brenner (1976, 1982) and Hilton (1981) saw feudalism, contrary to North and Thomas's (1973) model, as a set of institutions designed to extract rents from the peasants at the expense of social welfare.[9] Dependency theorists such as Williams (1944), Wallerstein (1974-1982), Rodney (1972), Frank (1978) and Cardoso and Faletto (1979) argued that

[9]Postan (1966, pp. 603-604) famously estimated that lords extracted about 50% of the entire production of peasants.

the international trading system was designed to extract rents from developing countries to the benefit of developed countries.

The social conflict view includes situations where economic institutions may initially be efficient for a set of circumstances but are no longer efficient once the environment changes. For example, Acemoglu, Aghion and Zilibotti (2001) show that though certain sorts of organizations may be useful for countries a long way from the technological frontier, it may be socially efficient to change them subsequently. This may not happen however because it is not privately rational. An interesting example may be the large business enterprises (the *chaebol*) of South Korea. In the context of political institutions, one might then develop a similar thesis. Certain sets of institutions are efficient for very poor countries but they continue to apply even after they cease to be the efficient institutional arrangement.

In stark contrast to the efficient institutions view, political institutions play a crucial role in the social conflict view. Which economic institutions arise depends on who has political power to create or block different economic institutions. Since political institutions play a central role in the allocation of such power they will be an intimate part of a social conflict theory of economic institutions.

What distinguishes the social conflict view from the ideological view is that social conflict can lead to choices of economic institutions which cause underdevelopment even when all agents have common knowledge that this is so. What distinguishes it from the incidental view is that it emphasizes that institutional choices which cause underdevelopment are conscious choices, rather than the result of some historical accident. The aspect that distinguishes the social conflict view from the efficient institutions view is that it does not assume that institutions are always efficient. This is one possible outcome but it is not the only one or indeed the most likely. Why is this? Why cannot efficiency be separated from distribution? We discuss this issue in the next section.

6 Sources of Inefficiencies

Having motivated our first two assertions in section 1.2, we are now in a position to discuss the third, related to the importance of commitment problems. The inability to commit to how political power will be used in the future means that the impact of economic institutions on efficiency cannot be separated from their effects on distribution.[10]

In any market situation where economic exchange takes place, and the quid is sepa-

[10]An alternative approach would be to stress informational asymmetries (Farrell, 1987).

rated from the pro quo, issues of commitment will arise. That these issues are of crucial importance has been recognized in the literatures on incomplete contracts and renegotiation (e.g., Hart, 1995). Nevertheless, if the legal system functions properly, there is an array of enforceable contracts that owners can sign with managers, workers with employers, borrowers with lenders etc. These contracts can be enforced because there is an authority, a third party, with the power to enforce contracts. Although the authority that is delegated to enforce contracts and to resolve disputes varies depending on the exact situation, all such power ultimately emanates from the state, which, in modern society, has a near-monopoly on the use of legitimate coercion. An owner and manager can write a contract because they believe that the state, and its agents the courts, would be impartial enforcers of the contract.

In contrast, if, for example, a manager believed that the state would be aligned with the interests of the owner and refuse to punish the owner if and when he failed to make a payment stipulated by the contract, then the contract would have little value. Therefore, the presence of an impartial enforcer is important for contracting. The problem when it comes to institutional choices is that there is no such impartial third party that can be trusted to enforce contracts. This is the origin of the commitment problem in politics.

To elaborate on this point, let us consider a situation where society can be governed as a dictatorship or as a democracy. Imagine that the dictator does not relinquish his power, but instead he promises that he will obey the rules of democracy, so that individuals can undertake the same investments as they would in democracy. This promise would not necessarily be credible. As long as the political system remains a dictatorship, there is no higher authority to make the dictator stick to his promise. There is no equivalent of a contract that can be enforced by an impartial third-party. After all, the dictator has the monopoly of military and political power, so he is the final arbiter of conflicting interests. There is no other authority to force the dictator to abide by his promises.

A similar problem plagues the reverse solution, whereby the dictator agrees to a voluntary transition to democracy in return for some transfers in the future to compensate him for the lost income and privileges. Those who will benefit from a transition to democracy would be willing to make such promises, but once the dictator relinquishes his political power, there is no guarantee that citizens would agree to tax themselves in order to make payments to this former dictator. Promises of compensation to a former dictator are typically not credible.

The essence of the problem is commitment. Neither party can commit to compensate the other nor can they commit to take actions that would not be in their interests ex

post. The reason why commitment problems are severe in these examples is because we are dealing with political power. Different institutions are associated with different distributions of political power, and there is no outside impartial party with the will and the power to enforce agreements. In some cases, there may be self-enforcing promises that maintain an agreement. Acemoglu (2003a) discusses such possibilities, but in general, there are limits to such self-enforcing agreements, because they require the participants to be sufficiently patient, and when it comes to matters of political power, the future is uncertain enough that no party would behave in a highly patient manner.

Based on this reasoning, we can now discuss three different channels via which the presence of commitment problems will lead to the choice and persistence of inefficient institutions.

6.1 Holdup

Imagine a situation in which an individual or a group holds unconstrained political power. Also suppose that productive investments can be undertaken by a group of citizens or producers that are distinct from the "political elites", i.e., the current power holders. The producers will only undertake the productive investments if they expect to receive the benefits from their investments. Therefore, a set of economic institutions protecting their property rights are necessary for investment. Can the society opt for a set of economic institutions ensuring such secure property rights? The answer is often no (even assuming that "society" wants to do so).

The problem is that the political elites—those in control of political power—cannot commit to respect the property rights of the producers once the investment are undertaken. Naturally, ex ante, before investments are undertaken, they would like to promise secure property rights. But the fact that the monopoly of political power in their hands implies that they cannot commit to not hold-up producers once the investments are sunk.

This is an obvious parallel to the hold-up problem in the theory of the firm, where once one of the parties in a relationship has undertaken investments specific to the relationship, other parties can hold her up, and capture some of the returns from her investments. As in the theory of the firm, the prospect of hold-up discourages investment. But now the problem is much more severe, since it is not only investments that are specific to a relationship that are subject to hold-up, but all investments.

This is therefore an example of how inefficient economic institutions arise because of a monopoly of political power. Those with political power cannot commit not to use their

political power ex post, and this translates directly into a set of economic institutions that do not provide secure property rights to groups without political power. The consequence is clear: without such protection, productive investments are not undertaken, and opportunities for economic growth go unexploited.

The reason why these inefficient economic institutions persist (or may be the equilibrium institutions of the society) is related to commitment problems. Parallel to our above example of inducing the dictator to relinquish power, there are two ways to introduce secure property rights. First, in principle, political elites could promise to respect property rights. However, mere promises would not be credible, unless backed up by the political elites relinquishing power, and this would mean relinquishing their rents and privileges. Second, political elites can be bought off by the beneficiaries of a system of more secure property rights. This would typically be achieved by a promise of future payments. For example, after investments are undertaken and output is produced, a share can be given to the political elites. But, as pointed out above, there is another, reverse commitment problem here; the beneficiaries of the new regime cannot commit to making the promised payments to the previous political elites.

Many real world examples illustrate the commitment problems involved in limiting the use of political power. In practice, although buying off dictators and persuading them to leave power is difficult, there have been many attempts to do so, usually by trying to guarantee that they will not be persecuted subsequently. One way of doing this is to give them asylum in another country. Nevertheless, such attempts rarely succeed, most likely again because of commitment problems (the new regime cannot commit to abide by its promises). An illustrative example of this is the attempts by the Reagan administration to persuade Jean-Claude ('Baby Doc') Duvalier to relinquish power in Haiti in 1986. In the face of a popular uprising and rising social and economic chaos, the Reagan administration, via the intermediation of the Jamaican Prime Minster Edward Seaga, tried to persuade Duvalier to go into exile. He at first agreed and the White House announced his departure on January 30th, but the next day he changed his mind, unsure that he would really be protected, and stayed in Haiti. One month later he was forced into exile in France by the military.

A more common, and in many ways more interesting strategy to induce dictators to relinquish power is to try to structure political institutions so as to guarantee that they will not be punished. Such institutional changes are sometimes important in transitions to democracy. For example, President Pinochet was willing to abide by the results of the 1989 plebiscite he lost in Chile because as a senator the Constitution protected him

from prosecution. It was only when he left the country that he was vulnerable.

Although Pinochet's experience illustrates an example of structuring political institutions to achieve commitment, to create durable institutions constraining future use of political power is difficult in practise. These difficulties are well illustrated by the transition from white rule in Rhodesia to majority rule in Zimbabwe. Facing an unwinable guerilla war, the white elite in Rhodesia sought to negotiate a transition of majority rule, but with enough institutional safeguards that their rents would be protected. These safeguards included the electoral system they wanted, which was used for the first post-independence elections, and massive over-representation in parliament (Reynolds 1999, p. 163). Whites were guaranteed 20% of the seats in the legislature for seven years despite making up only 2-3% of the population and were guaranteed 10 seats of the 40 seat senate. Clauses of the 1980 Constitution were also aimed at directly guaranteeing the property rights of the whites. In particular land reform was outlawed for 10 years after which it could only take place if compensated.

The white negotiators at the Lancaster House talks in 1979 that produced these agreements understood that any promises made by the black majority negotiators about what would happen after independence could not be believed. They sought therefore to find a set of rules that would get around this problem (Herbst, 1990, pp. 13-36). Nevertheless, these guarantees were not enough to protect the property rights (and rents) of the whites in anything other than the short run. The Mugabe regime quickly absorbed the other factions from among the African guerilla opposition, and more moderate relatively pro-white groups, such as Abel Muzorewa's United African National Council, crumbled. In 1985 the Mugabe regime switched back to the electoral system it preferred (Reynolds, 1999, p. 164) and in 1987, at the first possible opportunity, it removed the guaranteed representation for whites. Though in 1987 Mugabe nominated white candidates for these seats (Horowitz, 1991, pp. 135-136), this did not last for long. In 1990 the senate was abolished. Finally, in 1990 the Constitution was amended to allow for the redistribution of land. Since this time the Mugabe government has begun a sustained policy of land redistribution away from whites through legal and extra-legal means.

6.2 Political losers

Another related source of inefficient economic institutions arises from the desire of political elites to protect their political power. Political power is the source of the incomes, rents, and privileges of the elite. If their political power were eroded, their rents would decline. Consequently, the political elite should evaluate every potential economic change

not only according to its economic consequences, such as its effects on economic growth and income distribution, but also according to its political consequences. Any economic change that will erode the elites' political power is likely to reduce their economic rents in the long run.

As an example, imagine a change in economic institutions that will increase economic growth, but in doing so, will also enrich groups that could potentially contest political power in the future. Everything else equal, greater economic growth is good for those holding political power. It will create greater returns on the assets that they possess, and also greater incomes that they can tax or expropriate. However, if their potential enemies are enriched, this also means greater threats against their power in the future. Fearing these potential threats to their political power, the elites may oppose changes in economic institutions that would stimulate economic growth.

That the threat of becoming a political loser impedes the adoption of better institutions is again due to a commitment problem. If those who gained political power from institutional change could promise to compensate those who lost power then there would be no incentive to block better institutions.

There are many historical examples illustrating how the fear of losing political power has led various groups of political and economic elites to oppose institutional change and also introduction of new technologies. Perhaps the best documented examples come from the attitude of the elites to industrialization during the nineteenth century (see Acemoglu and Robinson, 2000b, 2002). There were large differences between the rates at which countries caught up with British industrialization with many countries completely failing to take advantage of the new technologies and opportunities. In most of these cases, the attitudes of political elites towards industrialization, new technology and institutional change appear to have been the decisive factor, and these attitudes were driven by their fears of becoming political losers. These issues are best illustrated by the experiences of Russia and Austria-Hungary.

In both Russia and Austria-Hungary, absolutist monarchies feared that promoting industrialization would undermine their political power. In Russia, during the reign of Nikolai I between 1825 and 1855 only one railway line was built in Russia, and this was simply to allow the court to travel between Moscow and St. Petersburg. Economic growth and the set of institutions that would have facilitated it were opposed since, as Mosse (1992) puts it "it was understood that industrial development might lead to social and political change." In a similar vein, Gregory (1991) argues: "Prior to the about face in the 1850's, the Russian state feared that industrialization and modernization would

concentrate revolution minded workers in cities, railways would give them mobility, and education would create opposition to the monarchy."

It was only after the defeat in the Crimean War that Nikolai's successor, Alexsandr II, initiated a large scale project of railway building and an attempt to modernize the economy by introducing a western legal system, decentralizing government, and ending feudalism by freeing the serfs. This period of industrialization witnessed heightened political tensions, consistent with the fears of the elites that times of rapid change would destabilize the political status quo and strengthen their opposition (McDaniel, 1988, gives a detailed account of these events, see also Mosse, 1958).

The consensus view amongst historians also appears to be that the main explanation for the slow growth of Austria-Hungary in the nineteenth century was lack of technology adoption and institutional change, again driven by the opposition of the state to economic change. This view was proposed by Gerschenkron who argued that the state not only failed to promote industrialization, but rather, "economic progress began to be viewed with great suspicion and the railroads came to be regarded, not as welcome carriers of goods and persons, but as carriers of the dreaded revolution. Then the state clearly became an obstacle to the economic development of the country" (1970, p. 89). See also Gross (1973).

The analysis of Fruedenberger (1967, pp. 498-499) is similar. As with the Tsar, the Hapsburg emperors opposed the building of railways and infrastructure and there was no attempt to develop an effective educational system. Blum (1943) pointed to the pre-modern institutional inheritance as the major blockage to industrialization arguing (p. 26) that

> "these living forces of the traditional economic system were the greatest barrier to development. Their chief supporter was ... Emperor Francis. He knew that the advances in the techniques of production threatened the life of the old order of which he was so determined a protector. Because of his unique position as final arbiter of all proposals for change he could stem the flood for a time. Thus when plans for the construction of a steam railroad were put before him, he refused to give consent to their execution 'lest revolution might come into the country'."

6.3 Economic losers

A distinct but related source of inefficiency stems from the basic supposition of the social conflict view that different economic institutions imply different distributions of incomes. This implies that a move from a bad to a better set of economic institutions will make some people or groups worse off (and will not be Pareto improving). This in turn implies that such groups will have an incentive to block or impede such institutional changes even if they benefit the whole of society in some aggregate sense.

The idea that economic losers impede the choice of efficient economic institutions and economic policies is widespread in economics and was seen earliest in the literature on international trade. Even though free trade may be socially desirable, individuals invested in sectors in which an economy does not enjoy comparative advantage will lose economically from free trade. Since at least the work of Schattshneider (1935) the role of economic losers has been central in understanding why free trade is not adopted. In the context of development economics, this idea was first discussed by Kuznets (1968), developed at length by Olson (1982, 2000) and Mokyr (1990), and formalized by Krusell and Rios-Rull (1996) and Parente and Prescott (1999). Most of the examples discussed in the development literature on economic losers are about technological change—people with specific investments in obsolete technology try to block the introduction of better technology. The most celebrated example is the case of the Luddites, skilled weavers in early nineteenth century England who smashed new mechanized looms which threatened to lead to massive cuts in their wages (see Thomis, 1970, Randall, 1991). Scott (2000, p. 200) relates a similar example from modern Malaysia, "When, in 1976, combine harvesters began to make serious inroads into the wages of poor villagers, the entire region experienced a rash of machine-breaking and sabotage reminiscent of the 1830's in England."

That better economic institutions are blocked by individuals whose incomes are threatened by such change is again due to a problem of commitment. If those whose incomes rose when economic institutions changed could promise to compensate those whose incomes fell then there would be no incentive to block better economic institutions. Nevertheless, it is difficult to commit to such transfers. To consider again the example of the Luddites, the factory owners could have promised to pay the weavers high wages in the future even though their skills were redundant. Once the new technology was in place however, owners would have a clear incentive to fire the weavers and hire

much cheaper unskilled workers.[11]

Although the problem of economic losers is appealing at first sight, has received some attention in the economics literature, and fits into our framework by emphasizing the importance of commitment problems, we view it both theoretically and empirically less important than the holdup and the political loser problems. First, as pointed out in Acemoglu and Robinson (2000b), in theories emphasizing issues of economic losers, there are implicit assumptions about politics, which, when spelled out, imply that political concerns must be important whenever issues of economic losers are present. The idea of economic losers is that certain groups, fearing that they will lose their economic rents, prevent adoption of beneficial economic institutions or technologies. The assumption in this scenario is that these groups have the political power to block socially beneficial changes. But then, if they have the political power to block change, why wouldn't they allow the change to take place and then use their political power to redistribute some of the gains to themselves? The implicit assumption must therefore be that groups losing economically also experience a reduction in their political power, making it impossible for them to redistribute the gains to themselves after to change takes place. This reasoning therefore suggests that whether certain groups will lose economically or not is not as essential to their attitudes towards change as *whether their political power will be eroded*. Problems of political losers therefore seem much more important than problems of economic losers.

Possibly for this reason, advocates of the economic losers view have been unable to come up with any well documented examples where the economic losers hypothesis can actually explain first-order patterns of development. For instance, while it is true that the Luddites tried to break machines, they singularly failed to halt the progress of agricultural technology in nineteenth century Britain. The same is true for Malaysia in the 1970s, one of the fastest growing economies in the world at that time. Neither set of workers had sufficient political power to stop change. Indeed, when political powerful groups became economic losers, such as landowners in nineteenth century England who saw land prices and agricultural rents fall rapidly after 1870, they did nothing to block change because their political power allowed them to benefit from efficient economic institutions (Acemoglu and Robinson, 2002).

[11] One possible way round this problem would be for the owners, if they could afford it, to compensate the weavers in advance for their lower future wages. But this would raise the reverse commitment problem: the weavers would have an incentive to take the money and still break the machines – i.e., they could not commit to not blocking the innovations that would reduce their wages even after they had taken the money.

Perhaps the most interesting failure of economic losers to halt progress in English economic history comes from the impact of the enclosure of common lands. Land has not always been privately owned as property. In much of Africa land is still owned communally, rather than individually, and this was true in Medieval Britain. Starting around 1550 however an 'enclosure movement' gathered pace where 'common land' was divided between cultivators and privatized. By 1850 this process of enclosures had made practically all of Britain private property.

Enclosure was a heterogenous process (Overton 1996, p. 147) and it also took place at different times in different places. Nevertheless, most of it was in two waves, the so called 'Tudor enclosures' between 1550 and 1700 and the 'parliamentary enclosures' in the century after 1750.

> "From the mid-eighteenth century the most usual way in which common rights were removed was through a specific act of parliament for the enclosure of a particular locality. Such acts ... made the process easier because enclosure could be secured provided the owners of a majority (four-fifths) of the land, the lord of the manor, and the owner of the tithe agreed it should take place. Thus the law of parliament (statue law) only took account of the wishes of those *owning* land as opposed to the common law which took account of all those who had both ownership rights and *use* rights to land. Moreover ... in some parishes the ... majority could be held by a single landowner ... parliamentary enclosure often resulted in a minority of owners imposing their will on the majority of farmers." Overton (1996, p. 158, italics in original)

The historical evidence is unanimous that the incentive to enclose was because "enclosed land was worth more than open common field land ... the general consensus has been that rents doubled" Overton 1996, p. 162). More controversial is the source of this increase in rent. Overton continues (pp. 162-163) "The proportion of profits taken as rents from tenants by landlords is the outcome of a power struggle between the two groups, and the increase in rent with enclosure may simply reflect an increase in landlord power." Allen (1982, 1992) showed, in his seminal study of the enclosure movement in the South Midlands, that the main impact was a large increase in agricultural rents and a redistribution of income away from those cultivators who had previously used the commons.

The enclosure of common land thus led to a huge increase in inequality in early modern England. Many peasants and rural dwellers had their traditional property rights expropriated. In protest, groups of citizens dispossessed by enclosure attempted to oppose it through collective action and riots—attempting to influence the exercise of political power. These groups were no match for the British state, however. Kett's rebellion of 1549, the Oxfordshire rebellion of 1596, the Midland Revolt of 1607, and others up to the Swing Riots of 1830-1831 were all defeated (see Charlesworth, 1983). The presence of economic losers did not prevent this huge change in economic institutions and income distribution.

6.4 The Inseparability of Efficiency and Distribution

Commitment problems in the use and the allocation of political power therefore introduce a basic trade-off between efficiency and distribution. For example, when lack of commitment causes hold-ups, those who hold political power know that people will not have the right incentives to invest so growth will be low. In response to this, they might voluntarily give away their power or try to create political institutions that restricted their power. Such a change in political institutions would create better investment incentives. Though this situation is hypothetically possible and has formed the basis for some theories of institutional change (e.g. Barzel, 2001) it appears to be insignificant in reality. Even faced with severe underinvestment, political elites are reluctant to give away their power because of its *distributional* implications, i.e., because this would reduce their ability to extract rents from the rest of society. Thus poor economic institutions, here lack of property rights and hold-up, persist in equilibrium because to solve the problem, holders of political power have to voluntarily constrain their power or give it away. This may increase the security of property in society and increase incentives to invest, but it also undermines the ability of rulers to extract rents. They may be better off with a large slice of a small pie.

Similar phenomena are at work when there are either political or economic losers. In the first case, namely a situation where political power holders anticipate being political losers, promoting good institutions directly reduces the political power and rents of incumbents and a similar trade-off emerges. Adopting efficient economic institutions will stimulate growth, but when the political status quo is simultaneously eroded the amount of rent accruing to the initially powerful may fall. In the second case, the incomes of those with political power to determine economic institutions falls directly when better economic institutions are introduced. In the absence of credible commitments to side-

payments, those whose incomes fall when better economic institutions are introduced have an incentive to block such institutions.

Because commitment problems seem so endemic in collective choice and politics, it seems natural to believe that institutional change has significant distributional consequences and as a result there will be conflict over the set of institutions in society.

6.5 Comparative Statics

Our analysis so far has made some progress towards our theory of differences in economic institutions. Although our full theory is yet to be developed in the later sections, the different mechanisms discussed in this section already point out the major comparative static implications of our approach regarding when economic institutions protecting the property rights of a broad cross-section of society are likely to be adopted, and when they are likely to be opposed and blocked. We now briefly discuss these comparative statics.

Hold-up, political loser and economic loser considerations lead to some interesting comparative static results which can be derived by considering the political institutions that lie behind these phenomena.

1. First, the perspective of hold-ups immediately suggests that situations in which there are constraints on the use of political power, for example, because there is a balance of political power in society or a form of separation of powers between different power-holders, are more likely to engender an environment protecting the property rights of a broad cross-section of society. When political elites cannot use their political power to expropriate the incomes and assets of others, even groups outside the elite may have relatively secure property rights. Therefore, constraints and checks on the use of political power by the elite are typically conducive to the emergence of better economic institutions

2. Second, a similar reasoning implies that economic institutions protecting the rights of a broad cross-section are more likely to arise when political power is in the hands of a relatively broad group containing those with access to the most important investment opportunities. When groups holding political power are narrower, they may protect their own property rights, and this might encourage their own investments, but the groups outside the political elites are less likely to receive adequate protection for their investments (see Acemoglu, 2003b).

3. Third, good economic institutions are more likely to arise and persist when there are only limited rents that power holders can extract from the rest of society, since such rents would encourage them to opt for a set of economic institutions that make the expropriation of others possible.

4. Finally, considerations related to issues of political losers suggest that institutional reforms that do not threaten the power of incumbents are more likely to succeed. Therefore, institutional changes that do not strengthen strong opposition groups or destabilize the political situation are more likely to be adopted.

6.6 The Colonial Experience in Light of the Comparative Statics

We now briefly return to the colonial experience, and discuss how the comparative statics discussed here shed light on the differences in economic institutions across the former colonies and the institutional reversal.

The second comparative static result above suggests a reason why better economic institutions developed in places where Europeans settled. In these societies, a relatively broad-based group of Europeans came to dominate political power, and they opted for a set of economic institutions protecting their own property rights. In contrast, in places where Europeans did not settle, especially where they were a small minority relative to a large indigenous population, they did not have the incentives to develop good economic institutions because such institutions would have made it considerably more difficult for them to extract resources from the rest of society.

The third comparative static suggests an important reason why in places with more wealth, resources and also a high density of indigenous population to be exploited, Europeans were more likely to opt for worse institutions, without any protection for the majority of the population, again because such institutions facilitated the extraction of resources by the Europeans.

The first comparative static result, in turn, is related to the persistence of the different types of economic institutions that Europeans established, or maintained, in different colonies. In colonies where Europeans settled in large numbers, they also developed political institutions placing effective checks on economic and political elites. In contrast, the political institutions in colonies with high population density, extractive systems of production, and few Europeans, concentrated power in the hands of the elite, and built a state apparatus designed to use coercion against the majority of the population. These different political institutions naturally implied different constraints on political and

economic elites. In the former set of colonies, there were constraints on the development of economic institutions that would favor a few at the expense of the majority. Such constraints were entirely absent in the latter set of colonies.

Finally, the fourth comparative static is useful in thinking about why many colonies did not attempt to change their economic institutions during the nineteenth century when new economic opportunities made their previous system based on forced labor, slavery, or tribute-taking much less beneficial relative to one encouraging investment in industry and commerce. Part of the answer appears to lie in the fact that the political power of the elites, for example of the plantation owners in the Caribbean, was intimately linked to the existing economic system. A change in the economic system would turn them into political losers, an outcome they very much wanted to avoid.

6.7 Reassessment of the Social Conflict View

So far we have shown that the econometric evidence is convincing that differences in economic institutions are the root cause of differences in prosperity. We then argued that although there are different approaches which can account for variation in economic institutions, the most plausible approach is the social conflict view. Though we believe that there are clear instances where history and ideology matter for the institutional structure of society, and clearly institutions are highly persistent, the most promising approach to understanding why different countries have different institutions is to focus on choices and their subsequent consequences. The social conflict view emphasizes the distributional implication of economic institutions and how commitment problems imply that efficiency and distribution cannot be separated. Hence the fundamental conflict within society over the nature of economic institutions has important implications for economic performance. Some economic institutions will promote growth, but they will not necessarily benefit all groups in society. Alternative economic institutions may induce economic stagnation, but may nevertheless enrich some groups. Which set of institutions results and whether or not a society prospers will be determined by which of these groups has the political power to get the institutions that differentially benefit them. At this point we have therefore substantiated the first three points we made in the introduction. To develop our theory of economic institutions further we need to be more specific about political power—where it comes from and why some people have it and not others. We undertake this task in section 8. Before doing this however the next section discusses three important historical examples of the evolution of economic institutions. We use these examples to show the explanatory power of the social conflict

view and to begin to illustrate in concrete settings how political power works.

7 The Social Conflict View in Action

We now discuss three important examples to bring out the fact that conflict over economic institutions is critical to the functioning of the economy and that this conflict stems, not from differences in beliefs, ideology or historical accidents, but from the impact of economic institutions on distribution. The examples also show that those with political power have a disproportionate effect on economic institutions and they show how the distribution of political power is influenced by different factors. These factors include the allocation of de jure political power through the structure of political institutions and the ability of groups to solve the collective action problem, or exercise what we called de facto political power. With these examples in mind in section 8 we move to discuss in more detail the nature and sources of political power.

7.1 Labor Markets

A market—an opportunity for individuals to exchange a commodity or service—is obviously a fundamental economic institution relevant for development. As Adam Smith (1776) argued, markets allow individuals to take advantage of the benefits of specialization and the division of labor, and scholars such as Pirenne (1937) and Hicks (1969) argued that the expansion of markets was perhaps *the* driving forces in long-run development.

In the history of Europe a key transformation was from feudal labor market institutions towards modern notions of a free labor market where individuals were able to decide who to work for and where to live. This process of institutional change was intimately connected to the transition from a whole set of feudal economic institutions to the economic institutions we think of as 'capitalist.' Most historians see this as key to the economic take-off that began in the nineteenth century. It was the countries which had made the transition away from feudalism most completely, such as England, the Netherlands and France, thanks to the revolution of 1789, which developed most rapidly. It was those where feudalism was still in operation, such as Russia and Austria-Hungary, which lagged far behind.

What can account for this differential evolution of feudalism? Scholars beginning with Postan (1937) saw the demographic collapse caused by the black death in the 1340's as demolishing feudalism in Western Europe. By dramatically altering the land/labor

ratio as approximately 40% of the population of Europe died (e.g., Cantor, 2001), the Black Death greatly increased the bargaining power of peasants and allowed them to negotiate a free status ending feudal obligations, particularly with respect to labor. Therefore, Postan's demographic theory implicitly emphasizes the role of political power in the decline of feudalism: this set of economic institutions started to disappear when the political power of the peasants increased and that of lords declined.

In fact, the distribution of power may be even more important in the whole story than Postan's theory suggests. As first pointed out by Brenner (1976), the demographic theory of the decline feudalism is not consistent with the comparative evidence. Although demographic trends were similar all over Europe and

> "it is true that ... in most of Western Europe serfdom was dead by the early sixteenth century. On the other hand, in Eastern Europe, in particular Pomerania, Brandenburg, East Prussia and Poland, decline in population from the late fourteenth century was accompanied by an ultimately successful movement towards imposing extra-economic controls, that is serfdom, over what had been, until then, one of Europe's freest peasantries. By 1500 the same Europe-wide trends had gone a long way towards establishing one of the great divides in European history, the emergence of an almost totally free peasant population in Western Europe, the debasement of the peasantry to unfreedom in Eastern Europe." (Brenner, 1976, p. 41).

What can explain these divergent outcomes? Brenner notes (p. 51): "It was the logic of the peasant to try to use his apparently improved bargaining position to get his freedom. It was the logic of the landlord to protect his position by reducing the peasants' freedom." The outcome "obviously came down to a question of power" (p. 51); whether the peasants or the lords had more political power determined whether serfdom declined or became stronger.

Although we are far from an understanding of the determinants of the relative structure of political power in different parts of Europe, Brenner suggests that an important element was the "patterns of the development of the contending agrarian classes and their relative strength in the different European societies: their relative levels of internal solidarity, their self-consciousness and organization, and their general political resources—especially their relationships to the non-agricultural classes (in particular, potential urban class allies) and to the state" (p. 52). To substantiate this view, Brenner studies how villages tended to be organized differently in Eastern Europe, there was

"more of a tendency to individualistic farming; less developed organization of collaborative agricultural practices at the level of the village or between villages; and little of the tradition of the 'struggle for commons rights' against the lords which was so characteristic of western development" (p. 57). This differential organization was due to the process of initial occupation of these Eastern lands.

Although many parts of Brenner's analysis remain controversial, there is general agreement that the decline of feudalism and the transformation of European labor markets were intimately related to the political power of the key groups with opposing interests, the peasants and the lords (see, for example, Aston and Philpin, 1985, on reactions to Brenner's interpretation). Feudal institutions, by restricting labor mobility and by removing the role of the labor market in allocating labor to jobs, undermined incentives and resulted in underdevelopment. But these same economic institutions created large rents for the aristocracy. As a consequence, aristocracies all over Europe attempted to maintain them. It was when their political power weakened that the process of transformation got underway.

7.2 Financial Markets

Much recent work on growth and development has focused on capital markets. Growth requires investment, so poor agents without access to financial markets will not have the resources to invest. Empirically many scholars have found correlations between the depth of financial markets and growth (see Levine, 2004) and absence of financial markets is at the heart of ambitious theories of comparative development by Banerjee and Newman (1993) and Galor and Zeira (1993).

If the stress on financial markets and financial intermediation is correct, a central issue is to understand why financial systems differ. For example, studies of the development of banking in the United States in the nineteenth century demonstrate a rapid expansion of financial intermediation which most scholars see as a crucial facilitator of the rapid growth and industrialization that the economy experienced. In his recent study Haber (2001, p. 9) found that in the United States, "In 1818 there were 338 banks in operation, with a total capital of $160 million-roughly three times as many banks and bank capital as in 1810. Circa 1860, the United States had 1,579 banks, with a total capital of $422.5 million. Circa 1914 there were 27,864 banks in the United States. Total bank assets totaled $27.3 billion."

One might see this rapid expansion of banking and financial services as a natural feature. Yet Haber (2001) shows that the situation was very different in Mexico (p. 24).

"Mexico had a series of segmented monopolies that were awarded to a group of insiders. The outcome, circa 1910 could not have been more different: the United States had roughly 25,000 banks and a highly competitive market structure; Mexico had 42 banks, two of which controlled 60 percent of total banking assets, and virtually none of which actually competed with another bank."

The explanation for this huge difference is not obvious. The relevant technology was certainly readily available everywhere and it is difficult to see why the various types of moral hazards or adverse selection issues connected with financial intermediation should have limited the expansion of banks in Mexico but not the United States. Haber then shows that (p. 9), "at the time that the U.S. Constitution was put into effect in 1789, ... [U.S. banking] was characterized by a series of segmented monopolies that shared rents with state governments via taxes or state ownership of bank stock. In some cases, banks also shared rents directly with the legislators who regulated them."

This structure, which looked remarkably like that which arose subsequently in Mexico, emerged because state governments had been stripped of revenues by the Constitution. In response, states started banks as a way to generate tax revenues. State governments restricted entry "in order to maximize the amount of rent earned by banks, rent which would then be shared with the state government in the form of dividends, stock distributions, or taxes of various types."

Thus in the early nineteenth century, U.S. banks evolved as monopolies with regulations aimed at maximizing revenues for the state governments. Yet this system did not last because states began competing among themselves for investment and migrants.

> "The pressure to hold population and business in the state was reinforced by a second, related, factor: the broadening of the suffrage. By the 1840s, most states had dropped all property and literacy requirements, and by 1850 virtually all states (with some minor exceptions) had done so. The broadening of the suffrage, however, served to undermine the political coalitions that supported restrictions on the number of bank charters. That is, it created a second source of political competition-competition within states over who would hold office and the policies they would enact."

The situation was very different in Mexico. After 50 years of endemic political instability the country unified under the highly centralized 40 year dictatorship of Porfirio Diaz until the Revolution in 1910.

In Haber's argument political institutions in the United States allocated political power to people who wanted access to credit and loans. As a result they forced state governments to allow free competitive entry into banking. In Mexico political institutions were very different. There were no competing federal states and the suffrage was highly restrictive. As a result the central government granted monopoly rights to banks who restricted credit to maximize profits. The granting of monopolies turned out to be a rational way for the government to raise revenue and redistribute rents to political supporters (see North, 1981, Chapter 3).

A priori, it is possible that the sort of market regulation Haber found in Mexico might have been socially desirable. Markets never function in a vacuum, but rather within sets of rules and regulations which help them to function. Yet it is hard to believe that this argument applies to Mexico (see also Maurer, 2002). Haber (2001) documents that market regulation was aimed not at solving market failures and it is precisely during this period that the huge economic gap between the United States and Mexico opened up (on which see Coatsworth, 1993, Engerman and Sokoloff, 1997). Indeed, Haber and Maurer (2004) examined in detail how the structure of banking influenced the Mexican textile industry between 1880 and 1913. They showed that only firms with personal contacts with banks were able to get loans. They conclude (p. 5):

> "Our analysis demonstrates that textile mills that were related to banks were less profitable and less technically efficient than their competitors. Nevertheless, access to bank credit allowed them to grow faster, become larger, and survive longer than their more productive competitors. The implication for growth is clear: relatively productive firms lost market share to relatively unproductive (but bank-related) competitors."

Despite the fact that economic efficiency was hurt by regulations, those with the political power were able to sustain these regulations.

7.3 Regulation of Prices

As our final example we turn to the regulation of prices in agricultural markets (which is intimately related to the set of agricultural policies adopted by governments). The seminal study of agricultural price regulation in Africa and Latin America is by Bates (1981, 1989, 1997). Bates (1981) demonstrated that poor agricultural performance in Ghana, Nigeria and Zambia was due to government controlled marketing boards systematically paying farmers prices for their crops much below world levels.

> "Most African states possess publicly sanctioned monopsonies for the purchase and export of agricultural goods ... These agencies, bequeathed to the governments of the independent states by their colonial predecessors, purchase cash crops for export at administratively determined domestic prices, and then sell them at the prevailing world market prices. By using their market power to keep the price paid to the farmer below the price set by the world market, they accumulate funds from the agricultural sector" Bates (1981, p. 12).

The marketing boards made surpluses which were given to the government as a form of taxation. Bates (1981, p. 15) notes

> "A major test of the intentions of the newly independent governments occurred ... [when] between 1959-1960 and 1961-62, the world price of cocoa fell approximately £50 a ton. If the resources generated by the marketing agencies were to be used to stabilize prices, then surely this was the time to use the funds for that purpose. Instead ... the governments of both Ghana and Nigeria passed on the full burden of the drop in price to the producers."

Bates continues "Using the price setting power of the monopsonistic marketing agencies, the states have therefore made the producers of cash crops a significant part of their tax base, and have taken resources from them without compensation in the form of interest payments or of goods and services returned." (pp. 181-9). As a result of this pernicious taxation, reaching up to 70% of the value of the crop in Ghana in the 1970's, investment in agriculture collapsed as did output of cocoa and other crops. In poor countries with comparative advantage in agriculture such a situation mapped into negative rates of economic growth.

Why were resources extracted in this way? Though part of the motivation was to promote industrialization, the main one is to generate resources that could be either expropriated or redistributed to maintain power

> "governments face a dilemma: urban unrest, which they cannot successfully eradicate through co-optation or repression, poses a serious challenge to their interests ... Their response has been to try to appease urban interests not by offering higher money wages but by advocating policies aimed at reducing the cost of living, and in particular the cost of food. Agricultural

policy thus becomes a by-product of political relations between governments and urban constituents" (1981, p. 33)

In contrast to the situation in Ghana, Zambia and Nigeria, Bates (1981, 1989, 1997) showed that agricultural policy in Kenya and Colombia over this period was much more pro-farmer. The difference was due to who controlled the marketing board. In Kenya, farmers were not smallholders, as they were in Ghana, Nigeria and Zambia, and concentrated landownership made it much easier to solve the collective action problem. Moreover, farming was important in the Kikuyu areas, an ethnic group closely related to the ruling political party, KANU, under Jomo Kenyatta (Bates, 1981, p. 122). Farmers in Kenya therefore formed a powerful lobby and were able to guarantee themselves high prices. Even though the government of Kenya engaged in land reform after independence

> "80% of the former white highlands were left intact and ... the government took elaborate measures to preserve the integrity of the large-scale farms ... [which] readily combine in defense of their interests. One of the most important collective efforts is the Kenya National Farmer's Union (KNFU) ... The organization ... is dominated by the large-scale farmers .. [but] it can be argued that the KNFU helps to create a framework of public policies that provides an economic environment favorable to all farmers" Bates (1981, pp. 93-94).

Bates concludes (p. 95) that in Kenya

> "large farmers ... have secured public policies that are highly favorable by comparison to those in other nations. Elsewhere the agrarian sector is better blessed by the relative absence of inequality. But is also deprived of the collective benefits which inequality, ironically, can bring."

In Colombia, farmers were favored because of competition for their votes from the two main political parties. Bates (1997, p. 54) notes

> "Being numerous and small, Colombia's coffee producers, like peasants elsewhere, encountered formidable costs of collective action. In most similar instances such difficulties have rendered smallholders politically powerless. And yet ... Colombia's peasants elicited favorable policies from politicians,

who at key moments themselves bore the costs of collective action, provisioning the coffee sector with economic institutions and delegating public power to coffee interests."

How could the coffee growers gain such leverage over national policy?

> "A major reason they could do so ... is because the structure of political institutions, and in particular the structure of party competition, rendered them pivotal, giving them the power over the political fortunes of those with ambition for office and enabling them to make or break governments. They thereby gained the power to defeat government officials who sought to orchestrate or constrain their behavior." Bates (1997, p.51, 54)

A telling piece of evidence in favor of this thesis is that during the 1950's when a civil war broke out between the two parties, there was five years of military rule and policy turned decisively again the coffee growers, only to switch back again with the peaceful resumption of democracy in 1958.

7.4 Political Power and Economic Institutions

These three examples of the creation of economic institutions have certain features in common. All these institutions, labor market regulation/feudalism, the rules governing financial market development, and agricultural price regulation, clearly reflect the outcome of conscious choices. Feudalism did not end in England for incidental or ideological reasons, but because those who were controlled and impoverished by feudal regulations struggled to abolish them. In Eastern Europe the same struggle took place but with a different outcome. Similarly, Mexico did not end up with different financial institutions than the United States by accident, because of different beliefs about what an efficient banking system looked like, or because of some historical factor independent of the outcome. The same is true for differences in economic policies in Kenya and Ghana. Moreover, different sets of economic institutions arising in different places cannot be argued to be efficient adaptations to different environments. Most historians believe that the persistence of feudal institutions in Eastern Europe well into the nineteenth century explains why it lagged far behind Western Europe in economic development. The difference between the financial institutions of Mexico and the United States also plausibly played a role in explaining why they diverged economically in the nineteenth century. The same holds with respect to agricultural price regulation.

The driving force behind all three examples is that economic institutions are chosen for their distributional consequences. Which specific economic institutions emerge depends on who is able to get their way—who has political power. In England, peasant communities had developed relatively strong local political institutions and were able to consolidate on the shock of the Black Death to put an end to feudal regulations. In Eastern Europe it was the lords who had relatively more power and they were able to intensify feudalism in the face of the same demographic shock (as Domar, 1970, pointed out, the Black Death actually made serfdom more attractive to the lords even if at the same time it increased the bargaining power of the peasants). In the case of banking in the nineteenth century, Haber's research shows while the authoritarian regime in Mexico had the political power to freely create monopolies and create rents in the banking industry, the United States was different because it was federal and much more democratic. The political institutions of the United States prevented politicians from appropriating the rents that could flow from the creation of monopolies. Finally, in Bates's analysis, distortionary price regulations arose in Ghana and Zambia, but not in Kenya and Colombia, because in the latter countries agricultural producers had more political power and so could prevent the distortionary policies that would harm their interests.

It is also useful to consider in the context of these examples the mechanisms we discussed in section 6 which underlie the adoption of inefficient economic institutions. Why couldn't the peasants and lords of feudal Europe negotiate and allow the introduction of a set of economic institutions that would have given peasants incentives to innovate and would have allowed for the efficient allocation of labor? Why couldn't either the lords have promised not to expropriate any benefits that accrued from innovation, or alternatively the peasants agreed to compensate the lords if feudal labor institutions were abolished? Though it is difficult to find direct evidence on such counterfactuals from the Medieval period, the most plausible explanation is that such 'deals' were impossible to make credible. The political power of the lords was intimately connected to feudal institutions and thus dismantling these would not only have increased peasant incentives to innovate, but would also have dramatically altered the balance of political power and the distribution of rents in society. Moreover, under feudal regulations peasants were tied to the land. The introduction of free labor mobility would have given workers an exit option, thus increasing their bargaining power with the lords over the division of output. Thus lords might anticipate being both political and economic losers from the ending of feudalism, even if total output would have increased.

In the case of agricultural price regulation, similar arguments are plausible. Cocoa farmers in Ghana would not have believed promises by governments that they would not expropriate the fruits of higher investment, and the governments themselves would not have believed promises by the farmers to compensate them if they left office. Moreover, efficient sets of economic institutions in Ghana or Nigeria would have strengthened the economic base of the rural sector at the expense of the political power of the then dominant urban sector. Indeed, for Ghana in the 1960's, we have direct evidence from the urban economy that the threat of being a political loser led to inefficient economic institutions. This emerges in the analysis of Killick (1978, p. 37) of the attempt by the government of Kwame Nkrumah to promote industrialization. Killick notes:

> "Even had there been the possibility [of creating an indigenous entrepreneurial class] it is doubtful that Nkrumah would have wanted to create such a class, for reasons of ideology and political power. He was very explicit about this saying 'we would be hampering our advance to socialism if we were to encourage the growth of Ghanian private capitalism in our midst.' There is evidence that he also feared the threat that a wealthy class of Ghanaian businessmen might pose to his own political power."

Further evidence on the importance of political loser considerations comes from E. Ayeh-Kumi one of Nkrumah's main economic advisers who noted after the coup that Nkrumah (Killick, 1978, p. 60): "informed me that if he permitted African business to grow, it will grow to becoming a rival power to his and the party's prestige, and he would do everything to stop it, which he actually did."

In this context, it is interesting that Nkrumah's solution to consolidate his power was to limit the size of businesses that Ghanaians could own. This caused problems for his industrialization policy which he got round by allowing foreign businessmen to enter Ghana. Though this was inconsistent with his aggressively nationalistic and anti-imperialistic rhetoric, these businessmen did not pose a domestic political threat. Killick (p. 37) notes "Given Nkrumah's desire to keep Ghanaian private businesses small, his argument that 'Capital investment must be sought from abroad since there is no bourgeois class amongst us to carry on the necessary investment' was disingenuous. He goes on to add that, (p. 40) Nkrumah "had no love of foreign capitalists but he preferred to encourage them rather than local entrepreneurs, whom he wished to restrict".

All these examples show that the distribution of political power in society is crucial for explaining when economic institutions are good and when they are bad. But where

does political power come from and who has political power? In addressing these questions we will develop our theory of economic institutions. In a theory based on social conflict where economic institutions are endogenous, it will be to differences in political institutions and the distribution of political power that we must look to explain variation in economic institutions.

8 A Theory of Institutions

8.1 Sources of Political Power

Who has political power and where does it come from? As we noted in the Introduction (section 1.2, point 4), political power comes from two sources. First, an individual or group can be allocated *de jure* power by *political institutions*. But institutions are not the only source of power. A second type of political power accrues to individuals or groups if they can solve the collective action problem, create riots, revolts, or demonstrations, own guns, etc.. We call this type of power *de facto* political power (see Acemoglu and Robinson, 2003, chapter 5).

Actual political power is the composition, the joint outcome, of de jure and de facto power. To see how this works out in practice, consider the situation in Chile in the early 1970's. Salvador Allende was elected President with a majority of the popular vote. The formal political institutions of democracy in Chile allocated power to him to propose legislation, issue decrees, etc. Consequently, even though he did not have an absolute majority in congress, Allende had a great deal of de jure political power. Political power is not just de jure however; it does not simply stem from political institutions. Allende, despite being empowered under the Chilean Constitution, was overthrown by a military coup in 1973. Here, the military under the leadership of General Pinochet, were able to use brute force and guns to over-ride the formal political institutions. The ability to use force is one example of de facto political power.

As we suggested in the introduction, the relationship between political power and economic and political institutions is complex and dynamic. Consider the example we discussed in section 7.2, the research by Haber on the comparative financial evolution of Mexico and the United States in the nineteenth century. Haber traced the different evolution of economic institutions to differences in initial political institutions. These political institutions led to different distributions of power and this was critical for the emergence of good financial institutions in the United States, whereby those who benefitted from a competitive banking industry were able to force politicians to provide the

rules which would guarantee it. But where did these differences in political institutions come from? These differences were partly a result of political events in the nineteenth century, and partially a result of different colonial political institutions. In the United States, during the initial phase of colonization in the early seventeenth century. Very low population density and lack of easily exploitable resources forced colonizing companies and the British state to make both economic and political concessions; they granted the settlers access to land and accepted the formation of representative democratic institutions (Morgan, 1975). Consequently, even at independence the United States had relatively democratic political institutions (Keyssar, 2000). Moreover, the initial egalitarian distribution of assets and the high degree of social mobility made for a situation where, at least in the northern states, the distribution of economic resources, and thus de facto power, was relatively equal. The relatively representative political institutions therefore persisted and were supported by the balance of de facto power in society.

In Mexico there were very different initial conditions during the colonial period with a large indigenous population and rich silver mines to exploit. This led to a much more hierarchical and authoritarian balance of political power and very different colonial economic institutions (see Engerman and Sokoloff, 1997, Acemoglu, Johnson and Robinson, 2004). These conditions fed into the different institutional structures at independence, the United States with its constitution, checks and balances and federalism, Mexico with its much more centralized, unchecked, unbalanced and absolutist state. These different political institutions then led to very different economic institutions and economic outcomes after independence. Thus, in some ultimate sense, the source of different political institutions were different initial conditions during the colonial period.

Consider now the evidence presented by Bates. Agricultural policies were better in Kenya because large farmers could solve the collective action problem and exercise de facto political power. But the main reason for the existence of large farms was that British settlers expropriated the land from Africans during the expansion of colonialism (see Berman and Lonsdale, 1992). Thus previous combinations of formal political institutions (colonial institutions) and de facto power (the military might of the British Empire) determined economic institutions, feeding into the future distribution of de facto power even after the nature of de jure power changed dramatically with independence.

We can now see that these examples substantiate the dynamic model that we sketched in section 1.2. There we showed that at any date, political power is shaped by political institutions, which determine de jure power, and the inherited distribution of resources, which affect the balance of de facto power. Political power then determines economic

institutions and economic performance. It also influences the future evolution of political power and prosperity. Economic institutions determine the distribution of resources at that point, which, in turn, influences the distribution of de facto power in the future. Similarly, the distribution of power at any point determines not just the economic institutions then, but also the future political institutions. Thus the allocation of political power at one date, because of the way it influences the distribution of resources and future political institutions, has a crucial effect on the future allocation of both de facto and de jure political power.

Both the comparison Haber made between Mexico and the United States, and that which Bates made between Ghana, Zambia, Kenya and Colombia illustrate this diagram in action. They show how political institutions and de facto power combine to generate different set of economic institutions, how these institutions determine both the distribution of resources and the growth rate of the economy, and how power and institutions evolve over time, often in ways that tend to reinforce particular initial conditions.

8.2 Political Power and Political Institutions

The examples we discussed above showed how political power depends on political institutions and de facto power, and how this determines economic institutions. Moreover, we saw that at any time the pre-existing economic institutions will be an important determinant of the distribution of de facto power. The final element to emphasize is how political institutions evolve over time and how they influence the distribution of political power.

To see why political institutions are so important as a source of political power think of a situation where a group, say the Chilean army in the early 1970's, has a great deal of de facto power. Indeed, it has so much de facto power that it can overrule the Chilean Constitution, making the political institutions largely irrelevant. In fact in Chile the de facto power of the military was able to overthrow the legitimate government and completely reverse the economic policies and economic institutions chosen by the Allende government (including land reform and mass nationalization of industry). Not only did the military reverse the economic institutions preferred by Allende and the groups who elected him, they then implemented their own preferred set of economic institutions, in particularly deregulating the trade regime and the economy. Yet the Pinochet regime was heavily concerned with formal political institutions, and in 1980 Pinochet re-wrote the constitution.

If de facto power was decisive in Chile what is the role for political institutions? If

the constitution can be overthrown, why bother to re-write it? The secret to this lies in the intrinsically transitory nature of de facto power.[12] Yes, the military were able to organize a coup in 1973 but this was only because times were uniquely propitious. There was a world-wide economic crisis, and factions of the military that opposed the coup could be marginalized. Moreover, the United States government at the time was happy to encourage and endorse the overthrow of a socialist government, even if it had been democratically elected. The coming together of such circumstances could not be expected to happen continually, hence once Chilean society re-democratized, as it did after 1990, the military would not be able to continually threaten a coup. In response to this Pinochet changed the political institutions in order to attempt to lock in the power of the military, and thus the economic institutions that he/they preferred. Therefore, the important role for political institutions is that they influence the future allocation of political power. This dynamic role is crucial because it explains the major desire of agents to change political institutions when they get the chance—this is how they can attempt to enduringly alter the balance of political power in their favor (see Acemoglu and Robinson, 2003).

8.3 A Theory of Political Institutions

We now have in place the outlines of our theory of institutions. There are seven points to emphasize, paralleling the discussion in section 1.2 and our diagrammatic exposition there. First, individuals have preferences over economic institutions because of the allocation of resources that these institutions induce.

Second, peoples' preferences typically do not agree because efficiency and distribution cannot be separated. Different economic institutions will benefit different groups, and this will determine the preferences of these individuals and groups with respect to economic institutions.

Third, the problem of commitment explains why efficiency and distribution are inseparable. Economic institutions are collective choices, and they are chosen and sustained

[12]The empirical literature on the collective action problem has recognized that the difficulty of solving the collective action problems lead collective action to typically be transitory. Lichbach (1995, p. 17) notes "collective action, if undertaken on a short-term basis, may indeed occur; collective action that requires long periods to time does not ... Given that most people's commitments to particular causes face inevitable decline, most dissident groups are ephemeral, most dissident campaigns brief." This transitory nature of collective action is echoed by Tarrow (1991, p. 15) who notes "the exhaustion of mass political involvement," while Ross and Gurr (1989, p. 414) discuss political "burnout." Similarly, Hardin (1995, p. 18) argues that "the extensive political participation of civil society receives enthusiastic expression only in moments of state collapse or great crisis. It cannot be maintained at a perpetually high level."

by the state. Since there is no third party to enforce the decisions of the state, problems of commitment are particularly severe in the political realm.

Fourth, the equilibrium structure of economic institutions will therefore be determined by who has the power to get their way, i.e., who can create and sustain economic institutions that benefit themselves. The distribution of political power thus determines economic institutions, the allocation of resources and the rate of economic growth.

Fifth, political power has two forms: de jure power determined by the political institutions, such as the constitution and the electoral rules, and de facto power, which stems from the ability to solve the collective action problem, mobilize weapons etc.. De facto power can influence political outcomes independently of the political institutions, and its distribution often critically determines how a given set of institutions works in practice and whether or not they are actually obeyed.

Sixth, the distribution of de facto political power at any date is influenced to a large degree by the distribution of resources in society, since those with greater resources can command more power both through legitimate and intimate means, and perhaps can also solve the collective action problem more efficiently. Naturally, the distribution of resources at this point is influenced by economic institutions and economic outcomes in the past.

Finally, political institutions are also endogenous; the current balance of political power, incorporating both de jure and de facto elements, also determines future political institutions. Political institutions are important because they allocate, at least within the limits defined by the exercise of future de facto power, the allocation of future de jure political power. Since de facto power, because of the nature of the collective action problem, is intrinsically transitory and difficult to wield, political institutions are often crucial in creating a source of durable political power. This makes it very attractive for groups to use their de facto political power to change political institutions so as to modify the distribution of future political power in their favor.

9 The Theory in Action

We now consider two examples that demonstrate our theory of institutions in action. Like the examples discussed in section 7, these examples contain all the elements of our theory laid out in a skeletal way in section 1.2. They show the role of political power in determining economic institutions, they demonstrate the different factors, both de facto and de jure, that determine political power, and they illustrate how de facto political power is often used to change political institutions in order to influenced the future

distribution of de jure political power.

9.1 Rise of Constitutional Monarchy and Economic Growth in Early Modern Europe

Our first example is the rise of constitutional monarchy in Europe. In the medieval period most European nations were governed by hereditary monarchies. However, as the feudal world changed, various groups struggled to gain political rights and reduce the autocratic powers of monarchies. In England, this process began as early as 1215 when King John was forced by his barons to sign the Magna Carta, a document which increased the powers of the barons, introduced the concept of equality before the law, and forced subsequent kings to consult with them. Many other European nations also developed 'parliaments' which kings could summon to discuss taxation or warfare (see Graves, 2001, Ertman, 1997). Nevertheless, the movement towards limited, constitutional monarchy was not linear or simple. Indeed, in France, certainly from the beginning of Louis XIV's reign in 1638, a more powerful absolutist monarchy appeared with very few controls. Indeed the feudal French parliaments, the Estates General, were not summoned between 1614 and 1788, just before the Revolution.

In England, the Tudor monarchs, particularly Henry VIII and then Elizabeth I, followed by the first Stuart kings, James I and Charles I, also attempted to build an absolutist monarchy. They failed, however, mostly because of Parliament, which blocked attempts to concentrate power. The constitutional outcome in England was settled by the Civil War from 1642-1651 and the Glorious Revolution in 1688. In the first of these conflicts the forces of Parliament defeated those loyal to Charles I and the king was beheaded. In 1660 the monarchy was restored when Charles II became king, but his brother James II was deposed in 1688 and Parliament invited William of Orange to become king.

Other places in Europe, particularly the Netherlands, saw similar developments to those in England. Under the Dukes of Burgundy, the Netherlands had won a considerable amount of political and economic freedom, particularly under the Grand Privilege of 1477 which gave the States General of the Burgundian Netherlands the right to gather on their own initiative and curbed the right of the ruler to raise taxes. However, the Netherlands were inherited by the Hapsburgs through marriage, and by 1493 Maximilian of Hapsburg had reversed the Grand Privilege. After 1552, war with France increased the Hapsburgs' fiscal needs and led them to impose a large tax burden on the Netherlands, already a prosperous agricultural and mercantile area. Growing fiscal and religious

resentment in 1572 led to a series of uprisings against the Hapsburgs, mostly orchestrated by commercial interests. These culminated in the War of Independence which was finally won in 1648.

While England and the Netherlands were developing limited constitutional governments, Spain and Portugal were moving in the same direction as France, towards greater absolutism. Davis (1973a, p. 66) notes [in Castille] "the king ruled subject only to weak constitutional restraints. In the first decades of the sixteenth century the crown had reduced the pretensions of the Castillian nobility and towns, so that the representative body, the Cortes, could obstruct but not in the last resort prevent royal tax raising."

These differential institutional trajectories were of enormous consequence. Netherlands and England moved ahead economically of the rest of Europe precisely because they developed limited, constitutional government. This form of government led to secure property rights, a favorable investment climate and had rapid multiplier effects on other economic institutions, particularly financial markets (see, e.g., North and Weingast, 1989, de Vries and van der Woude, 1997). While the Netherlands and Britain prospered, France was convulsed by the French Revolution, and by the nineteenth century Spain and Portugal were impoverished backward nations. How can we account for these diverging paths in the early modern period? Why did England and the Netherlands develop limited constitutional rule, while France, Spain and Portugal did not?

We proposed an explanation in Acemoglu, Johnson and Robinson (2002b) related to the differential responses of these countries to the opportunities of 'Atlantic trade', that is, overseas trade and colonial activity unleashed by the discovery of the New World and the rounding of the Cape of Good Hope at the end of the fifteenth century. All five nations engaged in Atlantic trade, but they did so in different ways, with very different implications for the organization of society, political institutions and subsequent economic growth.

In England "most trade was carried on by individuals and small partnerships, and not by the Company of Merchant Adventurers, the Levant Company ... or others of their kind" (Davis, 1973b, p. 41). At least by 1600 there was quite free entry into the English merchant class. The same was true in the Netherlands. In contrast, Cameron (1993, p. 127) describes the Portuguese situation as follows: "The spice trade in the East Indies of the Portuguese Empire was a crown monopoly; the Portuguese navy doubled as a merchant fleet, and all spices had to be sold through the *Casa da India* (India House) in Lisbon ... no commerce existed between Portugal and the East except that organized and controlled by the state." In Spain, similarly, colonial trade was a monopoly of the Crown

of Castille, which they delegated to the *Casa de Contratación* (House of Trade) in Seville. This merchants guild was closely monitored by the government (Parry, 1966, Ch 2). The main aim of these regulations was to make sure that all of the gold and silver from the Americas flowed back to Spain, creating a source of direct tax revenues for the crown. As a result, Latin American colonies were forbidden to buy manufactured goods from anywhere other than Spain, and all exports and imports had to pass through controlled channels. For example, until the Bourbon reforms of the mid eighteenth century, nothing could be exported directly from Buenos Aires, and if somebody produced anything for export on the Pampas, it had to be carried over the Andes and exported from Lima in Peru!

The source of the differences in the organization of trade, in turn, reflected the different political institutions of these countries. At the time, the granting of trade monopolies was a key fiscal instrument to raise revenues; the more powerful monarchs could increase their revenues by granting trade monopolies or by directly controlling overseas trade, while weaker monarchs could not. At the turn of the fifteenth century, the crown was much stronger in France, Spain and Portugal than in Britain and the Netherlands, and this was the most important factor in the differences in the organization of overseas trade. In fact, when both Tudor and Stuart monarchs attempted to create monopolies similar to those in Spain and Portugal, this was successfully blocked by the English Parliament (see, for example, Hill, 1969). Consequently, as world trade expanded in the sixteenth and early seventeenth centuries, it enriched merchants engaged in overseas trade in England and the Netherlands, but the crown and groups allied with it in France, Spain and Portugal. In England and the Netherlands, but not in France, Spain and Portugal, a new class of merchants (and gentry in England) arose with interests directly opposed to those of the Stuarts and the Hapsburgs, and this group was to play a central part in subsequent political changes.

In the case of the Netherlands, de Vries and van der Woude (1997) argue that "urban economic interests ultimately believed it advantageous to escape the Hapsburg imperial framework" (p. 369), and that it was "the traditional pillars of the maritime economy ... that supported and strengthened the young Republic in its hour of need" (p. 366). Moreover, in the case of Amsterdam, "[Hapsburgs'] opponents included most of the city's international merchants ... [I]n 1578 a new Amsterdam city council threw the city's lot in with the Prince of Orange ... among the merchants returning from ... exile were [those whose families] and several generations of their descendents would long dominate the city" (1997, p. 365). The expansion of world trade enriched and

expanded precisely those groups within Dutch society most opposed to Hapsburg rule. Israel (1995, pp. 241-242) writes: "From 1590, there was a dramatic improvement in the Republic's economic circumstances. Commerce and shipping expanded enormously, as did the towns. As a result, the financial power of the states rapidly grew, and it was possible to improve the army vastly, both qualitatively, and quantitatively, within a short space of time. The army increased from 20,000 men in 1588 to 32,000 by 1595, and its artillery, methods of transportation, and training were transformed" (see also Israel, 1989, Chapter 3). By 1629, the Dutch were able to field an army of 77,000 men, 50% larger than the Spanish army of Flanders (Israel, 1995, p. 507). As a consequence of the Dutch revolt, the Netherlands developed a republican form of government closely attuned to mercantile interests. De Vries and van der Woude (1997, p. 587) describe the new political elite following the Dutch Revolt as: "6 to 8% of urban households with incomes in excess of 1,000 guilders per year. This was the *grote burgerij* from whom was drawn the political and commercial leadership of the country. Here we find, first and foremost, the merchants," and point out how merchants dominated the governments of Leiden, Rotterdam and the cities in two largest states, Zeeland and Holland.

In England, the Civil War and Glorious Revolution coincided with the great expansion of English mercantile groups into the Atlantic. The East India Company was founded in 1600 as the culmination of a series of efforts to develop trade routes with Asia. The 1620s saw the great expansion of tobacco cultivation in Virginia. This was shortly followed by the development of the highly profitable English sugar colonies in the Caribbean. Finally, in the 1650s the English began to take over the Atlantic slave trade. Both the Civil War and the Glorious Revolution were at root battles over the rights and prerogatives of the monarchy. In both cases new merchant interests predominantly sided with those in the gentry demanding restrictions on the powers of the monarchy in order to protect their property and commerce.

The majority of merchants trading with the Americas and in Asia supported the Parliament during the Civil War. Brunton and Pennington (1954, p. 62) also note "in the country as a whole there was probably a preponderance of Parliamentarian feeling among merchants." Detailed analyses of the initial members of the Long Parliament in 1640 show that a significant majority of merchants supported the Parliamentarian cause (see Brenner, 1973, 1993, Keeler, 1954, and Brunton and Pennington, 1954). Members of the Commons from the City of London (the main center of mercantile activity), as well as many non-London commercial constituencies, such as Southampton, Newcastle and Liverpool, supported the Parliament against the King. These men included both

professional merchants and aristocrats who invested in colonizing the Americas. These new merchants also provided the financial support needed by the Parliament in the difficult early days of the war. They became the customs farmers for the new regime and therefore advanced tens of thousands of pounds that were essential in building up the army (Brenner, 1973, p. 82).

Pincus (1998, 2001, 2002) further documents the critical role of mercantile interests in the Glorious Revolution. He concludes (2002, p. 34) "England's merchant community actively supported William's plan for invasion, and provided a key financial prop to the regime in the critical early months." He notes that James II favored the East India Company and granted various monopoly privileges, alienating the merchant class. Thus, "no wonder the merchant community poured money into William of Orange's coffers in 1688." (Pincus, 2002, pp. 32-33).

The changes in the distribution of political power, political institutions and thus economic institutions that took place in England and the Netherlands had no counterparts in countries with relatively absolutist institutions, like Spain and Portugal, where the crown was able to closely control the expansion of trade. In these countries it was the monarchy and groups allied with it that were the main beneficiaries of the early profits from Atlantic trade, and groups favoring political and economic change did not become strong enough to induce such change. As a result, only in the Netherlands and England did constitutional rule emerge, and only in these two countries were property rights secure. As a result it was these same two countries that prospered.

Why could the monarchies of Spain and Portugal not negotiate a more efficient set of institutions? Alternatively why did the Stuart monarchs in England have to be beheaded or forced from power before better economic institutions could emerge?

It seems quite clear that a change to a more efficient set of institutions in Spain and Portugal would not have been possible under the auspices of the absolutist state, and a reduction in the power of the state was certainly inimical to the interest of the crown. In the case of England, Hill (1961a) argues directly that the reason that the Tudor and Stuart monarchs were not in favor of efficient economic institutions is because they feared that this would undermine their political power. He notes:

> "in general the official attitude to industrial advance was hostile, or at best indifferent. It was suspicious of social change and social mobility, the rapid enrichment of capitalists, afraid of the fluctuations of the market and of unemployment, of vagabondage, and social unrest ... the Elizabethan codes aimed at stabilizing the existing class structure, the location of industry and

the flow of labor supply by granting privileges and by putting hindrances in the way of the mobility and the freedom of contract."

The account so far explains why a change in the balance of (de facto) political power in England and the Netherlands led to a set of economic institutions favoring the interests of merchants. But in fact much more happened during the seventeenth century; an entirely new set of political institutions, constitutional regimes, restricting the power of the monarchy, were introduced. The reason why the merchants and the gentry in England (and the merchants in the Netherlands) used their newfound powers for political reform illustrates the dynamics of political power emphasized by our theoretical framework.

For example in the case of England, although in 1688 the Parliament might have been strong, it could not be sure that this power would endure. Indeed, the ability to solve the collective action problem and wield de facto power is intrinsically transitory. For instance, the Parliament vanquished James II with the help of a Dutch army, after which they invited William of Orange to take the throne. But how could they anticipate whether or not William would try to assert the absolutist prerogatives that James II had demanded?

The way to make transitory power permanent is to embody it into the rules of the game which is exactly what the English Parliament did after 1688. The changes in institutions after 1688 had large and important effects. For instance, in the eighteenth century the English monarchy was able to borrow huge amounts of money because the fiscal control of Parliament guaranteed that it would not default (see Brewer, 1988, Stasavage, 2003). This borrowing has been seen as crucial to the success of the English war machine. Moreover, with the Parliament in control of fiscal policy, the crown was never able to raise money through arbitrary taxation and not able any more to grant monopoly rights in exchange for money—an issue which had previously been a constant source of friction between the English crown and Parliament. Similarly, after 1688, the greater security of property rights in England led to a huge expansion of financial institutions and markets (Neal, 1990), which, North and Weingast (1989) argue, laid the institutional foundations for the Industrial Revolution.

Of course the English crown was not without some residual power and might have attempted to mount a coup against the Parliament to change political institutions back in its favor. This certainly happened in some places, such as in France after 1849 when Louis Napoleon mounted a successful coup to restore absolutist privileges lost in 1848. Nevertheless, changes in political institutions altered the nature of the status quo in

significant ways, and therefore, influenced the future distribution of de jure political power. Political institutions are not cast stone, and they can change, but they still create a source of political power more durable than mere de facto power.

9.2 Summary

The emergence of constitutional rule in some societies of early modern Europe therefore provides a nice example of how economic institutions, which shape economic outcomes, are determined by political power, which is in turn determined by political institutions and the distribution of resources in society. The Netherlands and England prospered in this period because they had good economic institutions, particularly secure property rights and well developed financial markets. They had these economic institutions because their governments were controlled by groups with a strong vested interest in such economic institutions. These groups wielded political power because of the structure of political institutions, i.e., they received de jure power in the Netherlands after the Dutch Revolt and in England after the Civil War and Glorious Revolution.

Moving one step back, we see that political institutions allocated more de jure political power to commercial interests in England and the Netherlands than in France, Spain and Portugal because of major changes in political institutions during the 1600s. These changes took place because commercial interests in England and the Netherlands acquired significant de facto political power as a result of their improving economic fortunes, itself a consequence of the interaction of Atlantic trade and the organization of overseas trade in these countries. Crucially for our framework, these commercial interests used their de facto power to reform (or revolutionize) political institutions so as to acquire de jure political power and solidify their gains.

These events, therefore, illustrate the various elements of our theoretical framework. In particular, they show how it is useful to think of political institutions and the distribution of economic resources as the state variables of the dynamic system, which determine the distribution of political power, and via this channel, economic institutions and economic outcomes. Political institutions and the distribution of economic resources are, themselves, endogenous, determined by political power and economic institutions, as exemplified by the fact that the distribution of economic resources changed significantly during the sixteenth century as a result of the new economic opportunities presented by the rise of Atlantic trade, and these changes were crucially influenced by the existing economic institutions (the organization of overseas trade). Furthermore, the change in the balance of political power led to the changes in political institutions through the

English Civil War, the Glorious Revolution and the Dutch Revolt.

9.3 Rise of Electoral Democracy in Britain

Our second example, based on Acemoglu and Robinson (2000a, 2001, 2003), is the rise of mass democracy. In the early nineteenth century, European countries were run by small elites. Most had elected legislatures, often descendents of medieval parliaments, but the franchise was highly restricted to males with relatively large amounts of assets, incomes or wealth. However, as the century and the Industrial Revolution progressed, this political monopoly was challenged by the disenfranchised who engaged in collective action to force political change.

In response to these developments, the elites responded in three ways. First by using the military to repress the opposition, as in the responses to the revolutions of 1848. Second, by making concessions to buy off opposition—this is the standard explanation for the beginnings of the welfare state in Germany under Bismarck. Finally, if neither repression nor concessions were attractive or effective, elites expanded the franchise and gave political power to the previously disenfranchised—they created the precedents of modern democracy.

The history of the rise of democracy in Britain is in many ways representative of the experiences of many other European countries. The first important move towards democracy in Britain came with the First Reform Act of 1832. This act removed many of the worst inequities under the old electoral system, in particular the 'rotten boroughs' where several members of parliament were elected by very few voters. The 1832 reform also established the right to vote based uniformly on the basis of property and income. The reform was passed in the context of rising popular discontent at the existing political status quo in Britain.

By the 1820s the Industrial Revolution was well under way and the decade prior to 1832 saw continual rioting and popular unrest. Notable were the Luddite Riots from 1811-1816, the Spa Fields Riots of 1816, the Peterloo Massacre in 1819, and the Swing Riots of 1830 (see Stevenson, 1979, for an overview). Another catalyst for the reforms was the July revolution of 1830 in Paris. Much of this was led and orchestrated by the new middle-class groups who were being created by the spread of industry and the rapid expansion of the British economy. For example, under the pre-1832 system neither Manchester nor Sheffield had any members of the House of Commons.

There is little dissent amongst historians that the motive for the 1832 Reform was to avoid social disturbances (e.g., Lang, 1999, p. 36). The 1832 Reform Act increased

the total electorate from 492,700 to 806,000, which represented about 14.5% of the adult male population. Yet, the majority of British people (the remaining 23 million) could not vote, and the elite still had considerable scope for patronage, since 123 constituencies still contained less than 1,000 voters. There is also evidence of continued corruption and intimidation of voters until the Ballot Act of 1872 and the Corrupt and Illegal Practices Act of 1883. The Reform Act therefore did not create mass democracy, but rather was designed as a strategic concession. In presenting his electoral reform to the British Parliament in 1831, the Prime Minister Earl Grey was well aware that this was a measure necessary to prevent a likely revolution. He argued:

> "The Principal of my reform is to prevent the necessity for revolution ... reforming to preserve and not to overthrow." (quoted in Evans, 1983, p. 212).

Unsurprisingly therefore, the issue of parliamentary reform was still very much alive after 1832, and it was taken up centrally by the Chartist movement. But as Lee (1994, p. 137) notes "The House of Commons was largely hostile to reform because, at this stage, it saw no need for it." This had changed by 1867, largely due to a juxtaposition of factors, including the sharp business cycle downturn that caused significant economic hardship and the increased threat of violence. Also significant was the founding of the National Reform Union in 1864 and the Reform League in 1865, and the Hyde Park riots of July 1866 provided the most immediate catalyst.

Lang (1999, p. 75) sums up his discussion by saying "The Hyde Park affair, coupled with other violent outbursts, helped to underscore the idea that it would be better to keep the goodwill of the respectable workers than to alienate them." Reform was initially proposed by the Liberal Prime Minister Russell in 1866 but was defeated by the Conservatives and dissident MP's. As a result Russell's government fell, and the Conservatives formed a minority administration with Lord Derby as their leader in the House of Lords, and Disraeli in charge of the House of Commons. It was Disraeli who then constructed a coalition to pass the Second Reform Act in 1867. As a result of these reforms, the total electorate was expanded from 1.36 million to 2.48 million, and working class voters became the majority in all urban constituencies. The electorate was doubled again by the Third Reform Act of 1884, which extended the same voting regulations that already existed in the boroughs (urban constituencies) to the counties (electoral constituencies in the rural areas). The Redistribution Act of 1885 removed many remaining inequalities in the distribution of seats and from this point on Britain

only had single member electoral constituencies (previously many constituencies had elected two members—the two candidates who gained the most votes). After 1884 about 60% of adult males were enfranchised. Once again social disorder appears to have been an important factor behind the 1884 act.

In Britain, the Reform Acts of 1867-1884 were a turning point in the history of the British state. Economic institutions also began to change. In 1871 Gladstone reformed the civil service, opening it to public examination, making it meritocratic. Liberal and Conservative governments introduced a considerable amount of labor market legislation, fundamentally changing the nature of industrial relations in favor of workers. During 1906-1914, the Liberal Party, under the leadership of Asquith and Lloyd George, introduced the modern redistributive state into Britain, including health and unemployment insurance, government financed pensions, minimum wages, and a commitment to redistributive taxation. As a result of the fiscal changes, taxes as a proportion of National Product more than doubled in the 30 years following 1870, and then doubled again. In the meantime, the progressivity of the tax system also increased (Lindert, 2004). Finally, there is also a consensus amongst economic historians that inequality in Britain fell after the 1870's (see Lindert, 2000, 2004)

Meanwhile, the education system, which was either primarily for the elite or run by religious denominations during most of the nineteenth century, was opened up to the masses; the Education Act of 1870 committed the government to the systematic provision of universal education for the first time, and this was made free in 1891. The school leaving age was set at 11 in 1893, then in 1899, it increased to 12 and special provisions for the children of needy families were introduced (Mitch, 1993). As a result of these changes, the proportion of 10-year olds enrolled in school that stood at 40 percent in 1870 increased to 100 percent in 1900 (Ringer, 1979, p. 207). Finally, a further act in 1902 led to a large expansion in the resources for schools and introduced the grammar schools which subsequently became the foundation of secondary education in Britain.

Following the Great War, the Representation of the People Act of 1918 gave the vote to all adult males over the age of 21, and women over the wage of 30 who were ratepayers or married to ratepayers. Ultimately, all women received the vote on the same terms as men in 1928. The measures of 1918 were negotiated during the war and may reflect to some extent a quid pro quo between the government and the working classes who were needed to fight and produce munitions. Nevertheless, Garrard (2002, p. 69) notes "most assumed that, if the system was to survive and 'contentment and stability prevail', universal citizenship could not be denied men, perceived to have suffered so

much and to have noticed Russia's Revolution."

Overall, the picture which emerges from British political history is clear. Beginning in 1832, when Britain was governed by the relatively rich, primarily rural aristocracy, a series of strategic concessions were made over an 86 year period to adult men. These concessions were aimed at incorporating the previously disenfranchised into politics since the alternative was seen to be social unrest, chaos and possibly revolution. The concessions were gradual because in 1832, social peace could be purchased by buying off the middle classes. Moreover, the effect of the concessions was diluted by the specific details of political institutions, particularly the continuing unrepresentative nature of the House of Lords. Although challenged during the 1832 reforms, the House of Lords provided an important bulwark for the wealthy against the potential of radical reforms emanating from a democratized House of Commons. Later, as the working classes reorganized through the Chartist movement and later through trade unions, further concessions had to be made. The Great War and the fallout from it sealed the final offer of full democracy. Though the pressure of the disenfranchised played less of a role in some reforms than others, and other factors undoubtedly played a role, the threat of social disorder was the main driving force behind the creation of democracy in Britain.

The story of the rise of mass democracy that emerges from the British evidence is one where economic and social changes connected with industrialization (for example, rising inequality) and urbanization increased the de facto power of the disenfranchised. In response, they demanded political rights, in particular changes in the political institutions which would allocate future political power to them. These changes in political institutions were, in many ways, the direct cause of the changes in economic institutions, in particular, in the labor market, in government policy, in the educational system, with major distributional implications, including the fall in inequality.

Why did elites in Britain create a democracy? Our discussion makes it clear that democracy did not emerge from the voluntary acts of an enlightened elite. Democracy was, in many ways, forced on the elite, because of the threat of revolution. Nevertheless, democratization was not the only potential outcome in the face of pressure from disenfranchised, or even in the face of the threat of revolution. Many other countries faced the same pressures and political elites decided to repress the disenfranchised rather than make concessions to them. This happened with regularity in Europe in the nineteenth century, though by the turn of the twentieth century most had accepted that democracy was inevitable. Repression lasted much longer as the favorite response of elites in Latin America, and it is still the preferred option for current political elites in China or Burma.

The problem with repression is that it is costly. Faced with demands for democracy political elites face a trade-off. If they grant democracy, then they lose power over policy and face the prospect of, possibly radical, redistribution. On the other hand, repression risks destroying assets and wealth. In the urbanized environment of nineteenth century Europe (Britain was 70% urbanized at the time of the Second Reform Act), the disenfranchised masses were relatively well organized and therefore difficult to repress. Moreover, industrialization had led to an economy based on physical, and increasing human, capital. Such assets are easily destroyed by repression and conflict, making repression an increasingly costly option for elites. In contrast, in predominantly agrarian societies like many parts of Latin America earlier in the century or current-day Burma, physical and human are relatively unimportant and repression is easier and cheaper. Moreover, not only is repression cheaper in such environments, democracy is potentially much worse for the elites because of the prospect of radical land reform. Since physical capital is much harder to redistribute, elites in Western Europe found the prospect of democracy much less threatening.

Faced with the threat of revolt and social chaos, political elites may also attempt to avoid giving away their political power by making concessions, such as income redistribution or other pro-poor policies. The problem with concessions however is their credibility, particularly when de facto power is transitory. For example, if a crisis, such as a harvest failure or business cycle recession creates a window of opportunity to solve the collective action problem and challenge the existing regime, the elites would like to respond with the promise of concessions. Yet windows of opportunity disappear and it is difficult to sustain collective action which entails people protesting in the streets and being away from their families and jobs. Thus collective action quickly dissipates and once it does so, the government has an incentive to renege on its promise of concessions. The promise of concessions, which people know to be non-credible is unlikely to defuse collective action. Hence, Acemoglu and Robinson (2000a, 2001, 2003) argue that democratization occurred as a way of making credible commitments to the disenfranchised. Democratization was a credible commitment to future redistribution, because it reallocated de jure political power away from the elites to the masses. In democracy, the poorer segments of the society would be more powerful and could vote, in other words, could use their de jure political power, to implement economic institutions and policies consistent with their interests. Therefore, democratization was a way of transforming the transitory de facto power of the disenfranchised poor into more durable de jure political power.

9.4 Summary

The emergence of mass democracy is another example illustrating our theory of institutions. Into the nineteenth century, economic institutions, particularly in the labor market, disadvantaged the poor. For example, trade unions were illegal and as late as the 1850 in Britain workers trying to organize a union could be shipped to the penal colony in Tasmania, Australia. The poor could not alter economic institutions in their favor because, being disenfranchised, they had little de jure political power and also limited de facto power,the because they were often unable to solve their collective action problems.

However, changes in the structure of society and the economy during the early nineteenth century altered the balance of political power, in particular making the exercise of de facto power by the politically disenfranchised much easier (Tilly, 1995, and Tarrow, 1998, document the changing qualitative nature of collective action over this period). The rise in the de facto political power of the poor necessitated a change in political institutions in their favor to defuse the threat of revolution. This was to tilt the future allocation of de jure political power, and consequently to ensure future economic institutions and policies consistent with their interests.

Whether or not increases in de facto power translated into democracy depended on a number of factors, in particular how difficult and costly it was for elites to use repression to counter the increase in the power of the masses, and how costly the prospect of democracy was. The changes in political institutions that occurred with democracy had profound implications for economic institutions. In the case of Britain, the period after the Second Reform Act of 1867 led the British state to commit itself to providing universal education for the first time and it also led to radical changes in labor market institutions allowing trade unions to form legally for the first time and increasing the bargaining power of labor. Hence economic institutions changed radically in favor of those newly endowed with de jure political power, mostly the relatively poor. This is in fact a relatively general result of democratization. Democracy enfranchises the poor, and the poor are able to use democracy to tilt economic institutions and the distribution of income in society in their favor (Li, Squire and Zou, 1998, Rodrik, 1999).

The emergence of democracy in the nineteenth-century Europe therefore also illustrates the workings of our theoretical framework. In particular, it shows how political institutions determine economic institutions and policies, and thus the distribution of resources, and it shows how political institutions change, especially in response to an im-

balance of de facto political power, as a credible way of influencing the future allocation of de jure political power.

10 Future Avenues

In this chapter we have proposed a framework for thinking about why some countries grow faster and are richer than others. We emphasized, following North and Thomas (1973), that most economic growth theory focuses only on proximate determinants. Although this body of work has been useful in helping us understand the mechanics of growth, it fails to provide a satisfactory account of why some countries grow while others do not. A major research goal must now be to get beyond the neoclassical growth model and its extensions, and search for the deeper causes, i.e., the fundamental determinants of growth.

We argued that the available evidence is consistent with the view that whether or not a society grows depends on how its economy is organized—on its economic institutions. We then proposed the outlines of a theory of institutions and illustrated it through a series of historical examples. We emphasized that a theory of why different countries have different economic institutions must be based on politics, on the structure of political power, and the nature of political institutions. Much remains to be done. First, the framework we outlined was largely verbal rather than mathematical, and thus, by its very nature, not fully specified. Constructing formal models incorporating and extending these ideas is the most important task ahead. Although some of our past work (e.g., Acemoglu and Robinson, 2000a, 2001, Acemoglu 2003b) formalizes parts of this framework, the full model has not been developed yet.

There are also many important issues left out of our framework, which appear to offer fruitful areas for future research. First, though we know that institutions, both economic and political, persist for long periods of time, often centuries (and sometimes millennia), we do not as yet have a satisfactory understanding of the mechanisms through which institutions persist.

Second, and closely related, although institutions do generally persist, sometimes they change. We have important examples of societies which have radically changed their political and economic institutions. Some do so for internal reasons, such as France after the Revolution of 1789, and some do because of external pressures such as Japan after the Meiji restoration or Russia after the Crimean War.

The important point here is that both institutional persistence and institutional change are equilibrium outcomes. Approaches positing institutional persistence as a

matter of fact, and then thinking of institutional changes as unusual events will not be satisfactory. Both phenomena have to be analyzed as part of the same dynamic equilibrium framework.

One type of institutional change, consistent with the examples we discussed in this chapter, takes place when those who benefit from the existing set of institutions are forced to accept change, either because they are the losers in a process of fighting or because of the threat of internal revolution (another possibility is that they might accept change because of the threat of external invasion). However, institutional change can also take place because the set of economic institutions that is optimal for a particular group with political power may vary over time as the state variables in the system and economic opportunities evolve. One example may be the end of slavery in the British Empire and another may be the economic and political changes introduced by Mikhail Gorbachev in the Soviet Union in the 1980s. We need more research on the dynamic mechanisms at work.

Finally, it is important to understand the role of policy and interventions in changing the institutional equilibrium. Though social science research is of intrinsic interest, one would hope that a convincing fundamental theory of comparative growth based on institutions would lead to policy conclusions that would help us improve the institutions and thus the lives and welfare of people in poor countries. It should be obvious that, at the moment, we are a long way from being in a position to draw such conclusions. In a world where political choices are made rationally and are endogenous to the structure of institutions, which are themselves ultimately endogenous, giving policy advice is a conceptually complex issue (see Acemoglu, Johnson, Robinson and Thaicharoen, 2003, for reflections on this issue). Recognizing our current ignorance on this topic in no way diminishes its importance, and its role as the Holy Grail of political economy research, however. And we believe that better and empirically more realistic theoretical frameworks in the future will take us closer to this Holy Grail.

11 References

Acemoglu, Daron (1995) "Reward Structures and the Allocation of Talent," *European Economic Review*, 39, 17-33.

Acemoglu, Daron (1997) "Training and Innovation in an Imperfect Labor Market," *Review of Economic Studies*, 64, 445-464.

Acemoglu, Daron (2003a) "Why Not a Political Coase Theorem?" NBER Working Paper #9377, forthcoming in the *Journal of Comparative Economics*.

Acemoglu, Daron (2003b) "The Form of Property Rights: Oligarchic versus Democratic Societies," NBER Working Paper #10037.

Acemoglu, Daron, Philippe Aghion and Fabrizio Zilibotti (2002) "Distance to Frontier, Selection, and Economic Growth," NBER Working Paper #9066.

Acemoglu, Daron, Simon Johnson and James A. Robinson (2001) "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, December, 91, 5, 1369-1401.

Acemoglu, Daron, Simon Johnson and James A. Robinson (2002a) "Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution," *Quarterly Journal of Economics*, 118, 1231-1294.

Acemoglu, Daron, Simon Johnson and James A. Robinson (2002b) "The Rise of Europe: Atlantic Trade, Institutional Change and Economic Growth" NBER Working Paper #9378.

Acemoglu, Daron, Simon Johnson and James A. Robinson (2004) *Institutional Roots of Prosperity*, the 2004 Lionel Robbins Lectures to be published by MIT Press.

Acemoglu, Daron, Simon Johnson, James A. Robinson and Yunyong Thaicharoen (2003) "Institutional Causes, Macroeconomic Symptoms: Volatility, Crises and Growth," *Journal of Monetary Economics*, 50, 49-123.

Acemoglu, Daron and James A. Robinson (2000a) "Why Did the West Extend the Franchise? Democracy, Inequality and Growth in Historical Perspective," *Quarterly Journal of Economics*, 115, 1167-1199.

Acemoglu, Daron and James A. Robinson (2000b) "Political Losers as a Barrier to Economic Development," *American Economic Review*, 90, 126-130.

Acemoglu, Daron and James A. Robinson (2001) "A Theory of Political Transitions," *American Economic Review*, 91, 938-963.

Acemoglu, Daron and James A. Robinson (2002) "Economic Backwardness

in Political Perspective," NBER Working Paper #8831

Acemoglu, Daron and James A. Robinson (2003) *Economic Origins of Dictatorship and Democracy*, Unpublished Book Manuscript.

Allen, Robert C. (1982) "The Efficiency and Distributional Consequences of Eighteenth Century Enclosures," *Economic Journal*, 92, 937-953.

Allen, Robert C. (1992) *Enclosure and the Yeoman*, New York; Oxford University Press.

Aston, T. H. and C. H. E. Philpin eds. (1985) *The Brenner debate: agrarian class structure and economic development in pre-industrial Europe*, New York; Cambridge University Press.

Bairoch, Paul (1988) *Cities and Economic Development: From the Dawn of History to the Present*, University of Chicago Press, Chicago.

Bairoch, Paul (1995) *Economics and World History: Myths and Paradoxes*, University of Chicago Press, Chicago.

Bairoch, Paul, Jean Batou and Pierre Chèvre (1988) *La Population des villes Europeenees de 800 a 1850: Banque de Données et Analyse Sommaire des Résultats*, Centre d'histoire économique Internationale de l'Uni. de Genève, Libraire Droz, Geneva.

Banerjee, Abhijit and Andrew F. Newman (1993) "Occupational Choice and the Process of Development," *Journal of Political Economy*, 101, 274-298.

Banfield, Edward C. (1958) *The moral basis of a backward society*, Chicago; University of Chicago Press.

Barro, Robert J. (1997) *The Determinants of Economic Growth: A Cross-Country Empirical Study*, Cambridge; MIT Press.

Barro, Robert J. and Rachel McCleary (2003) "Religion and Economic Growth," NBER Working Paper #9682.

Barzel, Yoram (2001) *A Theory of the State : Economic Rights, Legal Rights, and the Scope of the State*, New York; Cambridge University Press.

Bates, Robert H. (1981) *Markets and States in Tropical Africa*, University of California Press, Berkeley CA.

Bates, Robert H. (1989) *Beyond the Miracle of the Market*, New York; Cambridge University Press.

Bates, Robert H. (1997) *Open Economy Politics*, Princeton; Princeton University Press.

Becker, Gary S. (1958) "Competition and Democracy," *Journal of Law and Economics*, 1, 105-109.

Berman, Bruce J. and John Lonsdale (1992) *Unhappy Valley*, London; James Currey.

Bloom, David E. and Jeffrey D. Sachs (1998) "Geography, Demography, and Economic Growth in Africa," *Brookings Papers on Economic Activity*, 1998:2, 207-295.

Blum, Jerome (1943) "Transportation and Industry in Austria, 1815-1848," *Journal of Modern History*, 15, 24-38.

Brenner, Robert (1973) "The Civil War Politics of London's Merchant Community, *Past and Present*, 58, 53-107.

Brenner, Robert (1976) "Agrarian Class Structure and Economic Development in Preindustrial Europe" *Past and Present*, 70, 30-75.

Brenner, Robert (1982) "Agrarian Roots of European Capitalism" *Past and Present*, 97, 16-113.

Brenner, Robert (1993) *Merchants and Revolution: Commercial Change, Political Conflict, and London's Overseas Traders, 1550-1653*, Princeton; Princeton University Press.

Brewer, John (1988) *The sinews of power: war, money, and the English state, 1688-1783*, Cambridge; Harvard University Press.

Brunton, Douglas and D.H. Pennington (1954) *Members of the Long Parliament*, London; Allen and Unwin.

Cameron, Rondo (1993) *A Concise Economic History of the World*, New York; Oxford University Press.

Cantor, Norman F. (2001) *In the wake of the plague: the Black Death and the world it made*, New York; The Free Press.

Cass, David (1965) "Optimum Growth in an Aggregate Model of Capital Accumulation," *Review of Economic Studies*, 32, 233-240.

Chandler, Tertius (1987) *Four Thousand Years of Urban Growth: An Historical Census*, St. David's University Press, Lewiston, N.Y.

Charlesworth, Andrew (1983) *An Atlas of rural protest in Britain 1548-1900*, Croon Helm; London.

Coatsworth, John H. (1993) "Notes on the Comparative Economic History of Latin America and the United States," in Walter L. Bernecker and Hans Werner Tobler eds. *Development and Underdevelopment in America: Contrasts in Economic Growth in North and Latin America in Historical Perspective*, New York; Walter de Gruyter.

Coase, Ronald H. (1937) "The Nature of the Firm," *Economica*, 3, 386-405.

Coase, Ronald H. (1960) "The Problem of Social Cost," *Journal of Law and*

*Economic*s, 3, 1-44.

Cumings, Bruce (2004) *North Korea: Another Country*, New York; The New Press.

Curtin, Philip D. (1989) *Death by Migration: Europe's Encounter with the Tropical World in the nineteenth Century*, New York; Cambridge University Press.

Curtin, Philip D. (1998) *Disease and Empire: The Health of European Troops in the Conquest of Africa*, New York; Cambridge University Press.

Davis, Ralph (1973a) *The Rise of the Atlantic Economies*, Ithaca; Cornell University Press.

Davis, Ralph (1973b) *English Overseas Trade 1500-1700*, Macmillan; London.

Demsetz, Harold (1967) "Toward a Theory of Property Rights," *American Economic Review*, 57, 61-70.

de Vries, Jan and Ad van der Woude (1997) *The First Modern Economy: Success, Failure, and Perseverance of the Dutch Economy, 1500-1815,* New York; Cambridge University Press.

Diamond, Jared M. (1997) *Guns, Germs and Steel: The Fate of Human Societies*, W.W. Norton & Co., New York NY.

Djankov, Simeon, Rafael LaPorta, Florencio Lopez-de-Silanes, Andrei Shleifer (2002) "The Regulation of Entry," *Quarterly Journal of Economics*, 117, 1-37.

Djankov, Simeon, Rafael LaPorta, Florencio Lopez-de-Silanes, Andrei Shleifer (2003) "Courts," *Quarterly Journal of Economics*, 118, 453-517.

Dobb, Maurice H. (1948) *Studies in the Development of Capitalism*, Cambridge University Press, Cambridge UK.

Domar, Evsey (1970) "The Causes of Slavery or Serfdom: A Hypothesis," *Journal of Economic History*, 30, 18-32.

Durlauf, Steven N. and Marcel Fafchamps (2003) "Empirical Studies of Social Capital: A Critical Survey," Unpublished, Department of Economics, University of Wisconsin at Madison, http://www.ssc.wisc.edu/econ/archive/wp2003-12.pdf.

Edgerton, Robert B.(1992) *Sick Societies: challenging the myth of primitive harmony*, New York; Free Press.

Eggimann, Gilbert (1999) *La Population des Villes des Tiers-Mondes, 1500-1950*, Centre d'histoire économique Internationale de l'Uni. de Genève, Libraire Droz, Geneva.

Engerman, Stanley L. and Kenneth L. Sokoloff (1997) "Factor Endowments, Institutions, and Differential Growth Paths among New World Economies," in Stephen

Haber ed. *How Latin America Fell Behind*, Stanford University Press, Stanford CA.

Ertman, Thomas (1997) *Birth of the leviathan: building states and regimes in medieval and early modern Europe*, New York; Cambridge University Press.

Evans, Eric J. (1983) *The Forging of the Modern State: Early Industrial Britain, 1783-1870*, Longman; New York.

Farrell, Joseph (1987) "Information and the Coase Theorem," *Journal of Economic Perspectives*, 1, 113-129.

Frank, Andre Gunder (1978) *Dependent Accumulation and Underdevelopment*, Macmillan, London.

Freudenberger, Herman (1967) "State Intervention as an Obstacle to Economic Growth in the Hapsburg Monarchy," *Journal of Economic History*, 27, 493-509.

Galor, Oded and Joseph Zeira (1993) "Income Distribution and Macroeconomics," *Review of Economic Studies*, 40, 35-52.

Garrard, John (2002) *Democratization in Britain: Elites, Civil Society and Reform since 1800*, Basingstoke; Palgrave.

Gerschenkron, Alexander (1970) *Europe in the Russian Mirror: Four Lectures in Economic History*, Cambridge University Press, Cambridge UK.

Glaeser, Edward L. and Andrei Shleifer (2002) "Legal Origins," *Quarterly Journal of Economics*, 117, 1193-1230.

Graves, Michael A.R. (2001) *The Parliaments of Early Modern Europe*, New York; Longman.

Gregory, Paul R. (1991) "The Role of the State in Promoting Economic Development: the Russian Case and its General Implications," in Richard Sylla and Gianni Toniolo eds, *Patterns of European Industrialization: The Nineteenth Century*, Routledge, New York.

Greif, Avner (1994) "Cultural Beliefs and the Organization of Society: A Historical and Theoretical Reflection on Collectivist and Individualist Societies," *Journal of Political Economy*, 102, 912-950.

Gross, Nachaum (1973) "The Industrial Revolution in the Hapsburg Monarchy, 1750-1914," in Carlo M. Cipolla ed. *The Fontana Economic History of Europe, Volume 4*, Fontana Books, London.

Grossman, Sanford J. and Oliver D. Hart (1986) "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy*, 94, 691-719.

Gutierrez, Hector (1986) "La Mortalite des Eveques Latino-Americains aux XVIIe

et XVIII Siecles," *Annales de Demographie Historique*, 29-39.

Haber, Stephen H. (2002) "Political Institutions and Banking Systems: Lessons from the Economic Histories of Mexico and the United States, 1790-1914," Unpublished, Department of Political Science, Stanford University.

Haber, Stephen H. and Noel Maurer (2004) "Related Lending and Economic Performance: Evidence from Mexico," Unpublished, Department of Political Science, Stanford University.

Hall, Robert E. and Charles I. Jones (1999) "Why Do Some Countries Produce so much more Output per Worker than Others?" *Quarterly Journal of Economics*, 114, 83-116.

Hardin, Russell (1995) *All For One*, Princeton University Press; Princeton.

Hart, Oliver D. (1995) *Firms, contracts, and financial structure*, New York; Oxford University Press.

Herbst, Jeffrey I. (1990) *State Politics in Zimbabwe*, Berkeley; University of California Press.

Hill, Christopher (1961a) *The Century of Revolution,1603-1714*, New York; W.W. Norton & Co.

Hill, Christopher (1961b) "Protestantism and the Rise of Capitalism," in F.J. Fisher ed. *Essays in the Economic and Social History of Tudor and Stuart England*, Cambridge University Press; Cambridge.

Hill, Christopher (1969) *From Reformation to Industrial Revolution 1530-1780*, Baltimore; Penguin Books.

Hilton, Rodney (1981) *Bond Men Made Free*, Routledge, Oxford.

Horowitz, Donald L. (1991) *A Democratic South Africa? Constitutional Engineering in Divided Societies*, Berkeley; University of California Press.

Israel, Jonathan I. (1989) *Dutch Primacy in World Trade, 1585-1740*, Oxford; The Clarendon Press.

Israel, Jonathan I. (1995) *The Dutch Republic: Its Rise, Greatness and Fall 1477-1806*, New York; Oxford University Press.

Jones, Eric L. (1981) *The European Miracle: Environments, Economies, and Geopolitics in the History of Europe and Asia*, Cambridge University Press, New York.

Keeler, Mary (1954) *The Long Parliament, 1640-1641; A Biographical Study of its Members*, Philadelphia; American Philosophical Society.

Keyssar, Alexander (2000) *The right to vote: the contested history of democracy in the United States*, New York; Basic Books.

Killick, Tony (1978) *Development Economics in Action: a study of economic policies in Ghana*, London; Heinemann.

Knack, Steven and Philip Keefer (1995) "Institutions and Economic Performance: Cross-Country Tests using Alternative Measures," *Economics and Politics*, 7, 207-227.

Knack, Steven and Philip Keefer (1997) "Does Social Capital have an Economic Impact? A Cross-Country Investigation," *Quarterly Journal of Economics*, 112, 1252-1288.

Koopmans, Tjalling C. (1965) "On the Concept of Optimal Economic Growth," in *The Economic Approach to Development Planning*, Amsterdam; North-Holland.

Krusell, Per and Jose-Victor Rios-Rull (1996) "Vested Interests in a Theory of Stagnation and Growth," *Review of Economic Studies*, 63, 301-330.

Kupperman, Karen O. (1993) *Providence Island, 1630-1641: The other Puritan Colony*, New York; Cambridge University Press.

Kuznets, Simon (1968) *Towards a Theory of Economic Growth*, Yale University Press, New Haven CT.

Landes, David S. (1998) *The Wealth and Poverty of Nations: Why Some Are So Rich and Some So Poor*, W.W. Norton & Co., New York.

Lang, Sean (1999) *Parliamentary Reform, 1785-1928*, New York; Routledge.

La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer and Robert Vishny (1998) "Law and Finance," *Journal of Political Economy*, 106, 1113-55.

La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer and Robert Vishny (1999) "The Quality of Government," *Journal of Law, Economics and Organization*, 15, 222-279.

Lee, Stephen J. (1994) *Aspects of British Political History, 1815-1914*, Routledge; New York.

Levine, Ross (2004) "Finance and Growth," Chapter 21 of Philippe Aghion and Steven N. Durlauf eds. the *Handbook of Economic Growth*, to be published by North-Holland.

Lewis, W. Arthur (1978) *The Emergence of the International Economic Order*, Princeton; Princeton University Press.

Li, Hongyi, Lyn Squire and Heng-fu Zou (1998) "Explaining International and Intertemporal Variations in Income Inequality," *Economic Journal*, 108, 26-43.

Lichbach, Mark I. (1995) *The Rebel's Dilemma*, University of Michigan Press; Ann Arbor.

Lindert, Peter H. (2000) "Three Centuries of Inequality in Britain and America," in Anthony B. Atkinson and François Bourguignon eds. *Handbook of Income Distribution*, North-Holland, Amsterdam.

Lindert, Peter H. (2004) *Growing Public: Social Spending and Economics Growth since the Eighteenth Century*, Two volumes. Cambridge University Press, 2004.

Lucas, Robert E. (1988) "On the Mechanics of Economic Development," *Journal of Monetary Economics*, 22, 3-42.

McDaniel, Timothy (1991) *Autocracy, Modernization and Revolution in Russia and Iran*, Princeton; Princeton University Press.

McEvedy, Colin and Richard Jones (1978) *Atlas of World Population History*, New York; Facts on File.

Maddison, Angus (2001) *The World Economy: A Millenial Perspective*, Development Centre of the Organization for Economic Cooperation and Development, OECD, Paris.

Marshall, Alfred [1890] (1949) *Principles of Economics*, London; Macmillan.

Maurer, Noel (2002) *The Power and the Profits: The Mexican Financial System, 1876-1932*, Stanford; Stanford University Press.

Mitch, David (1983) "The Role of Human Capital in the First Industrial Revolution," in Joel Mokyr ed. *The British Industrial Revolution: An Economic Perspective*, San Francisco; Westview Press.

Mokyr, Joel (1990) *The Lever of Riches: Technological Creativity and Economic Progress*, New York; Oxford University Press.

Montesquieu, Charles de Secondat [1748] (1989) *The Spirit of the Laws*, New York; Cambridge University Press.

Morgan, Edmund S. (1975) *American Slavery, American Freedom: the Ordeal of Colonial Virginia*, New York; W.W. Norton & Co.

Moore, Barrington Jr. (1966) *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World*, Boston; Beacon Press.

Mosse, W.E. (1958) *Alexsandr II and the Modernization of Russia*, University of London Press, London UK.

Mosse, W.E. (1992) *An Economic History of Russia, 1856-1914*, I.B. Taurus Press, London, UK.

Murphy, Kevin J., Andrei Shleifer and Robert W. Vishny (1989a) "Industrialization and the Big Push," *Journal of Political Economy*, 97, 1003-1026.

Murphy, Kevin J., Andrei Shleifer and Robert W. Vishny (1989b) "Income

Distribution, Market Size and Industrialization," *Quarterly Journal of Economics*, 104, 537-564.

Myrdal, Gunnar (1968) *Asian Drama; An Inquiry into the Poverty of Nations*, 3 Volumes, Twentieth Century Fund, New York.

Neal, Larry (1990) *The rise of financial capitalism: international capital markets in the age of reason*, New York; Cambridge University Press.

Newton, Arthur P. (1914) *The Colonizing Activities of the English Puritans,* New Haven; Yale University Press.

North, Douglass C. (1981) *Structure and Change in Economic History*, New York; W.W. Norton & Co.

North, Douglass C. (1990) *Institutions, Institutional change, and Economic Performance*, Cambridge University Press, New York.

North, Douglass C. and Robert P. Thomas (1973) *The Rise of the Western World: A New Economic History*, Cambridge University Press, Cambridge UK.

North, Douglass C., William Summerhill and Barry R. Weingast (2000) "Order, Disorder, and Economic Change: Latin America versus North America," in Bruce Bueno de Mesquita and Hilton L. Root eds. *Governing for Prosperity*, New Haven; Yale University Press.

North, Douglass C. and Barry R. Weingast (1989) "Constitutions and Commitment: Evolution of Institutions Governing Public Choice in Seventeenth Century England, *Journal of Economic History*, 49, 803-832.

Olson, Mancur (1982) *The Rise and Decline of Nations: Economic Growth, Stagflation, and Economic Rigidities,* Yale University Press, New Haven and London.

Olson, Mancur (2000) *Power and Prosperity: Outgrowing Communist and Capitalist Dictatorships,* Basic Books, New York.

Overton, Mark (1996) *Agricultural Revolution in England: The Transformation of the Agrarian Economy 1500-1850*, Cambridge University Press; New York.

Parente Stephen and Edward C. Prescott (1999) "Monopoly Rights as Barriers to Riches," *American Economic Review*, 89, 1216-1233.

Parry, James H. (1966) *The Spanish Seaborne Empire*, Berkeley; University of California Press.

Piketty, Thomas (1995) "Social Mobility and Redistributive Politics," *Quarterly Journal of Economics*, 100, 551-584.

Pincus, Steven (1998) "Neither Machiavellian Moment nor Possessive Individualism: Commercial Society an the Defenders of the English Commonwealth," *American*

Historical Review, 103, 705-736.

Pincus, Steven (2001) "From Holy Cause to Economic Interest: The Study of Population and the Invention of the State," in Alan Houston and Steven Pincus eds. *A Nation Transformed: England after the Restoration*, New York; Cambridge University Press.

Pincus, Steven (2002) "Civic Republicanism and Political Economy in an Age of Revolution: Law, Politics, and Economics in the Revolution of 1688-89," Unpublished, Department of History, University of Chicago.

Pirenne, Henri (1937) *Economic and Social History of Medieval Europe*, New York; Harcourt, Brace and Company.

Postan, M.M. (1937) "The Chronology of Labour Services," *Transactions of the Royal Historical Society*, 20,

Postan, M. M. (1966) "Medieval Agrarian Society in its Prime: England," in M.M. Postan ed. *The Cambridge Economic History of Europe*, London; Cambridge University Press.

Putnam, Robert D. (with Robert Leonardi and Raffaella Y. Nanetti) (1993) *Making democracy work: civic traditions in modern Italy*, Princeton; Princeton University Press.

Randall, Adrian (1991) *Before the Luddites: custom, community, and machinery in the English woollen industry, 1776-1809*, New York; Cambridge University Press.

Reynolds, Andrew (1999) *Electoral Systems and Democratization in Southern Africa*, New York; Oxford University Press.

Ringer, Fritz (1979) *Education and Society in Modern Europe*, Bloomington; University of Indiana Press.

Robinson, James A. (1998) "Theories of Bad Policy," *Journal of Policy Reform*, 3, 1-46.

Rodrik, Dani (1999) "Democracies Pay Higher Wages," *Quarterly Journal of Economics,* CXIV, 707-738.

Romer, David (2003) "Misconceptions and Political Outcomes," *Economic Journal*, 113, 1-20.

Romer, Paul M. (1986) "Increasing Returns and Long-Run Growth," *Journal of Political Economy*, 94, 1002-1037.

Romer, Paul M. (1990) "Endogenous Technical Change," *Journal of Political Economy*, 98, 71-102.

Rosenstein-Rodan, Paul (1943) "Problems of Industrialization in Eastern and

south-eastern Europe," *Economic Journal*, 53, 202-211.

Ross, Jeffrey I. and Tedd R. Gurr (1989) "Why Terrorism Subsides: A Comparative Study of Canada and the United States," *Comparative Politics*, 21, 405-426.

Sachs, Jeffrey D. (2000) "Notes on a New Sociology of Economic Development" in Lawrence E. Harrison and Samuel P. Huntington eds. *Culture Matters: How Values Shape Human Progress*, Basic Books, New York.

Sachs, Jeffrey D. (2001) "Tropical Underdevelopment," NBER Working Paper #8119.

Schattschneider, Elmer E. (1935) *Politics, pressures and the tariff; a study of free private enterprise in pressure politics, as shown in the 1929-1930 revision of the tariff*, New York, Prentice-Hall, inc.

Scott, James C. (2000) "The Moral Economy as an Argument and as a Fight," in Adrian Randall and Andrew Charlesworth eds. *Moral Economy and Popular Protest: Crowds, Conflict and Authority*, London; MacMillan.

Smith, Adam [1776] (1999) *The Wealth of Nations* (Two Volumes), Penguin Classics, London.

Solow, Robert M. (1956) "A Contribution to the Theory of Economic Growth," *Quarterly Journal of Economics*, 70, 65-94.

Stasavage, David (2003) *Public Debt and the Birth of the Democratic State: France and Great Britain, 1688-1789*, New York; Cambridge University Press.

Stevenson, John (1979) *Popular Disturbances in England, 1700-1870*, New York; Longman.

Tarrow, Sidney (1991) "Aiming at a Moving Target: Social Science and the Recent Rebellions in Eastern Europe," *PS: Political Science and Politics*, 24, 12-20.

Tarrow, Sidney (1998) *Power in Movement: Social Movements and Contentious Politics*, Second Edition, New York; Cambridge University Press.

Tawney, R.H. (1926) *Religion and the Rise of Capitalism: A Historical Study*, London; J. Murray.

Tawney, R.H. (1941) "The Rise of the Gentry, 1558-1640," *Economic History Review*, 11, 1-38

Thomis, Malcolm I. (1970) *The Luddites; machine-breaking in regency England*, Newton Abbot; David & Charles.

Thompson, I.A.A. (1994) "Castile: Absolutism, Constitutionalism and Liberty," in Philip T. Hoffman and Kathryn Norberg eds. *Fiscal Crisis, Liberty and Representative Government*, Stanford; Stanford University Press.

Tilly, Charles (1995) *Popular Contention in Britain, 1758-1834*, Cambridge; Harvard University Press.

Townsend, Robert M. (1993) *The medieval village economy: a study of the Pareto mapping in general equilibrium models*, Princeton; Princeton University Press.

Vargas, Llosa, Mario (1989) *The storyteller*, New York; Farrar, Straus, Giroux.

Veitch, John M. (1986) "Repudiations and Confiscations by the Medieval State," *Journal of Economic History*, 46, 31-36.

Véliz, Claudio (1994) *The New World of the Gothic Fox: Culture and Economy in English and Spanish America*, Berkeley; University of California Press.

Wallerstein, Immanuel M. (1974-1980) *The Modern World-System*, 3 Volumes, Academic Press, New York.

Weber, Max (1930) *The Protestant Ethic and the Spirit of Capitalism*, Allen and Unwin; London.

Weber, Max (1958) *The Religion of India*, Free Press; Glencoe.

Wiarda, Howard J. (2001) *The Soul of Latin America: The Cultural and Political Tradition*, New Haven; Yale University Press.

Williams, Eric E. (1944) *Capitalism and Slavery*, University of North Carolina Press, Chapel Hill.

Williamson, Oliver (1985) *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*, New York; Free Press.

Wittman, Donald (1989) "Why Democracies Produce Efficient Results," *Journal of Political Economy*, 97, 1395-1424.

Figure 1
Average Protection Against Risk of Expropriation 1985-95 and log GDP per capita 1995

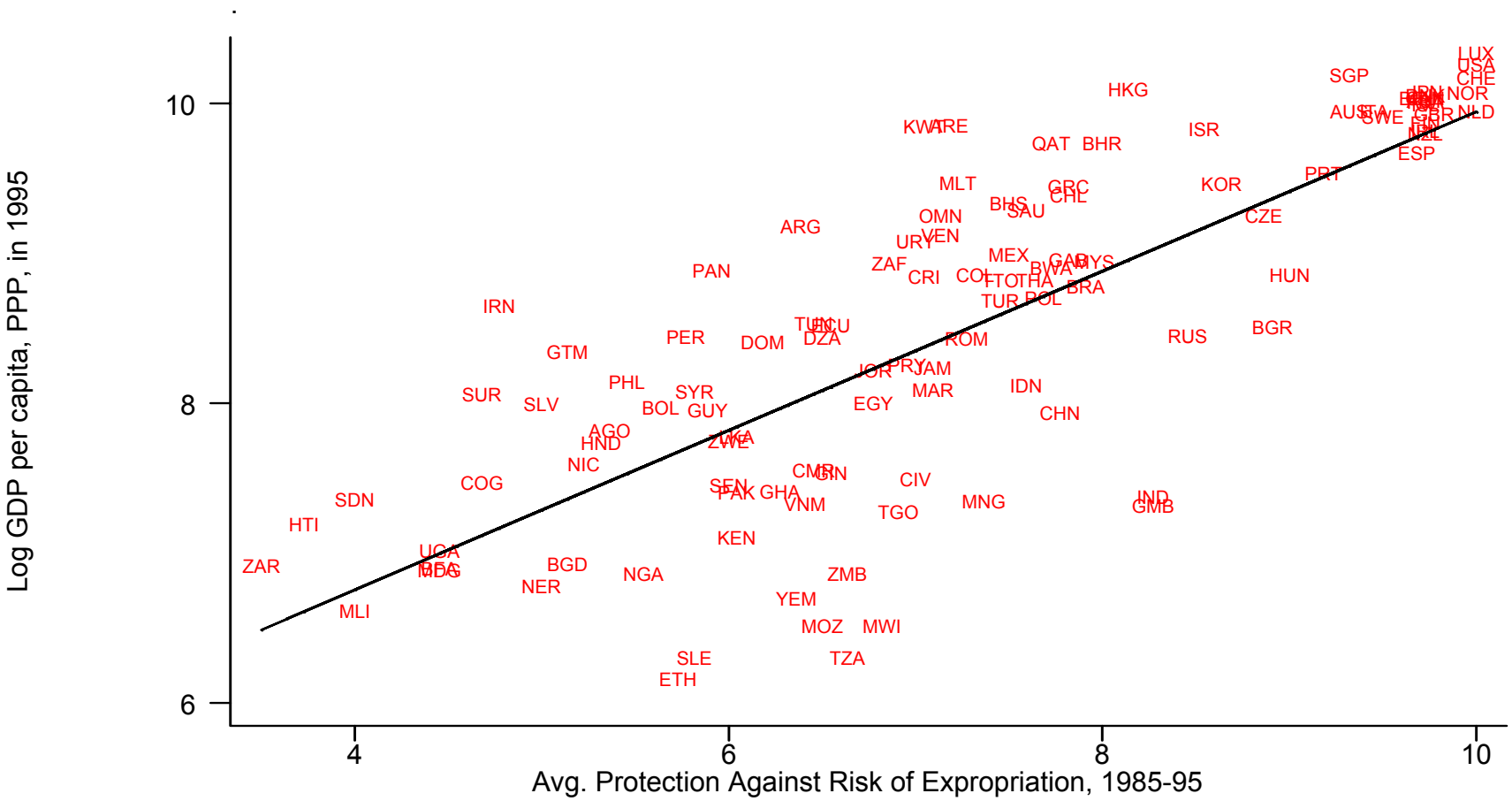


Figure 2
Latitude and log GDP per capita 1995

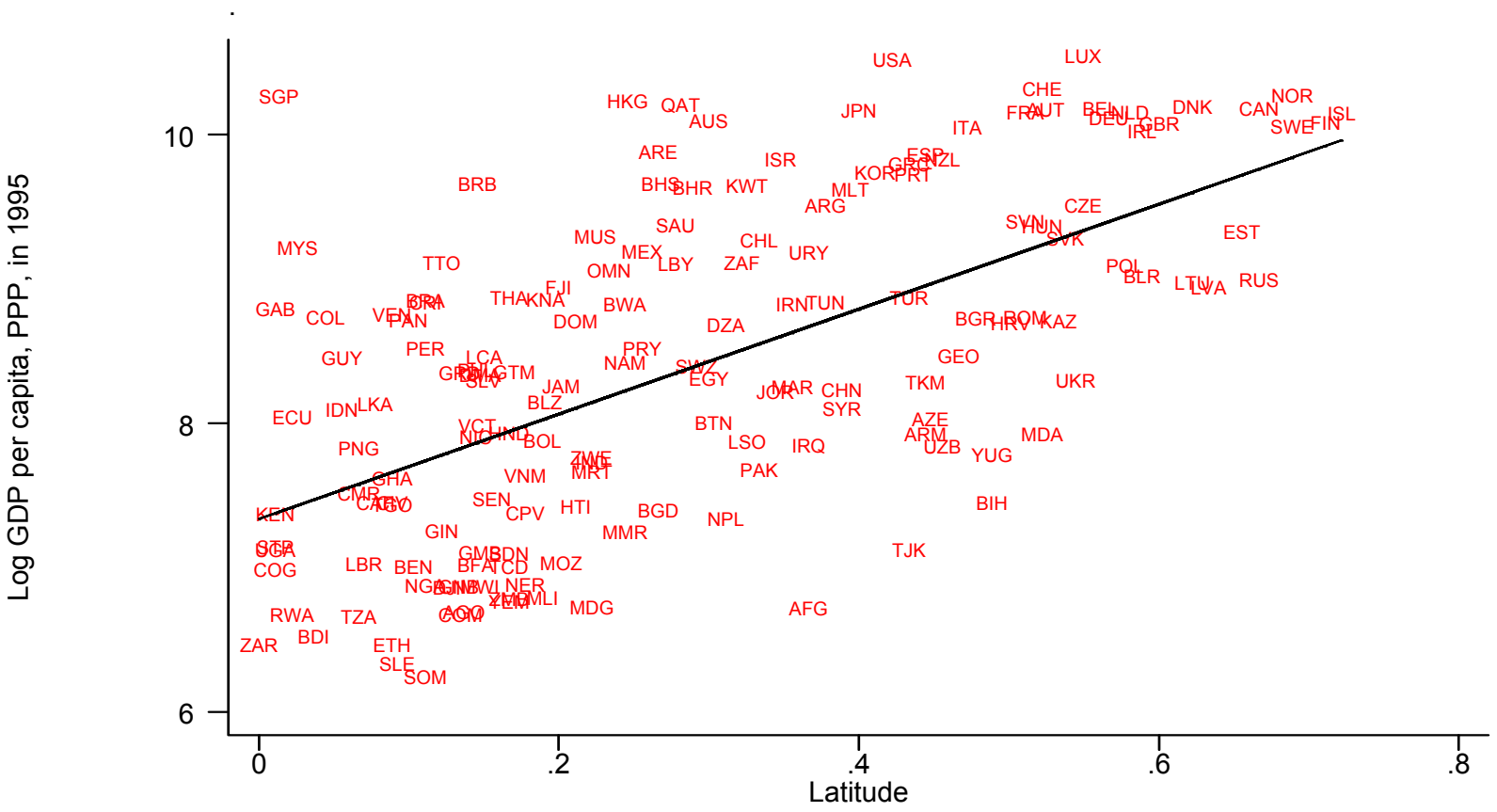


Figure 3
GDP per capita in North and South Korea, 1950-98

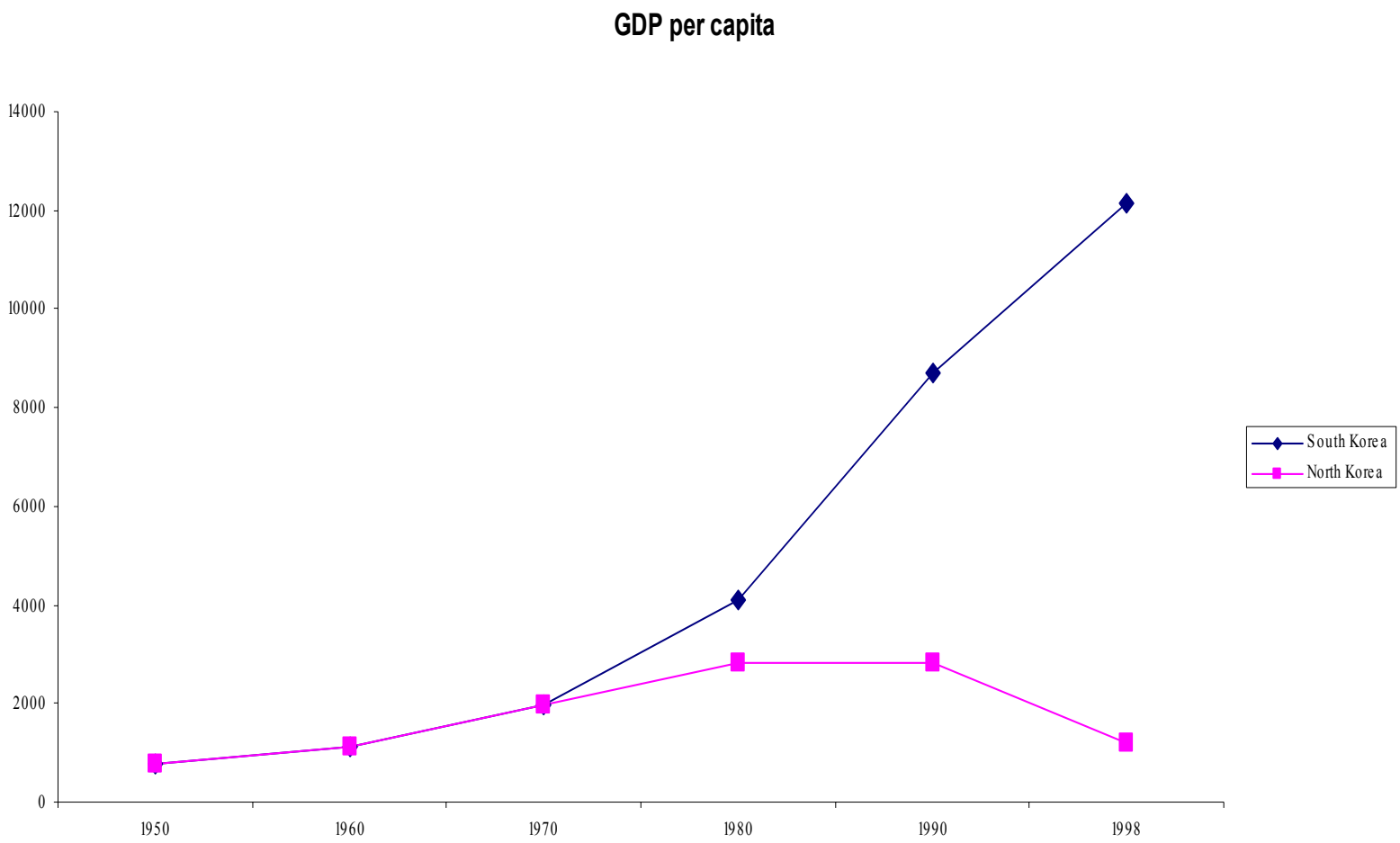


Figure 4
Urbanization in 1995 and log GDP per capita in 1995

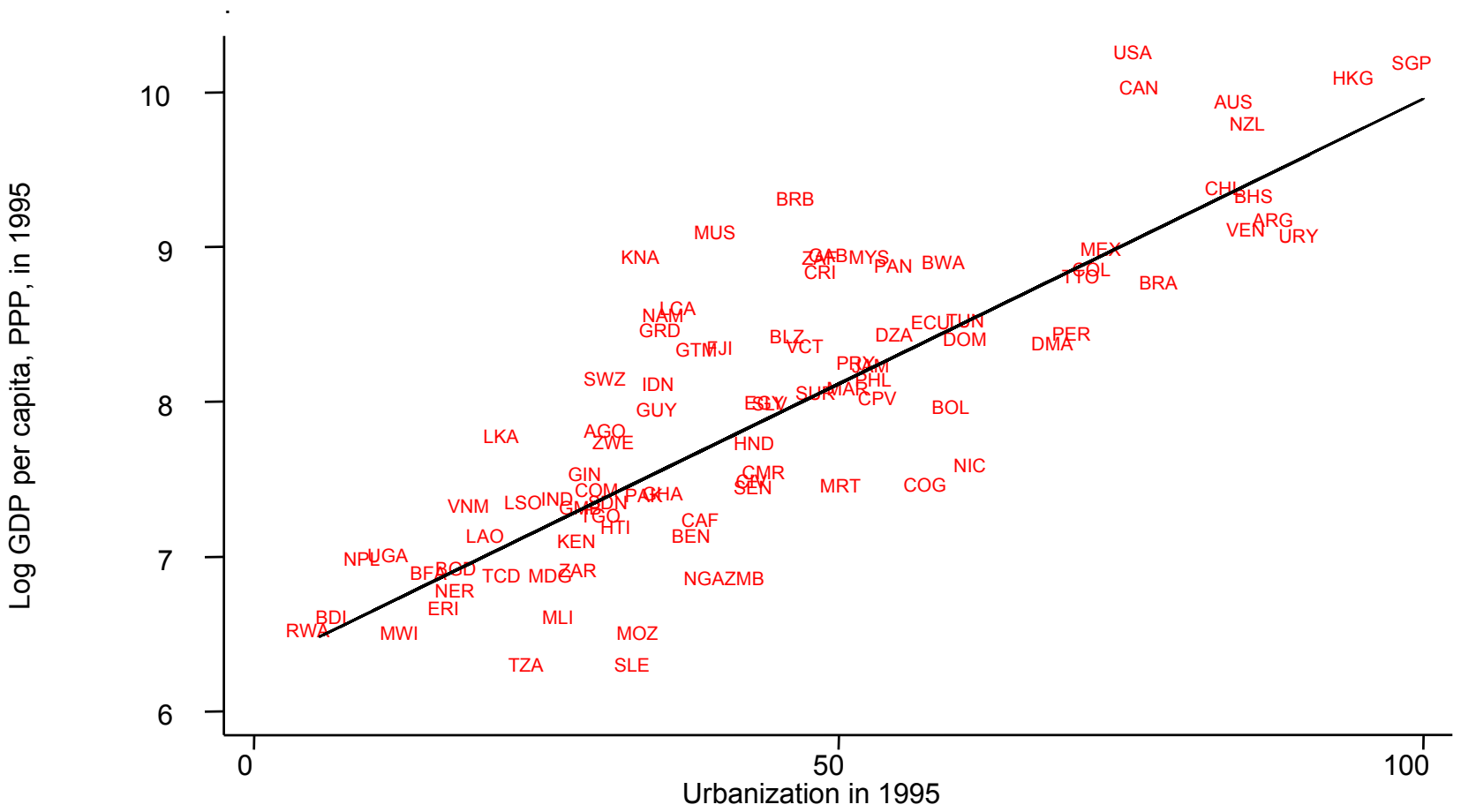


Figure 5
Urbanization in 1500 and log GDP per capita in 1995, among former European colonies

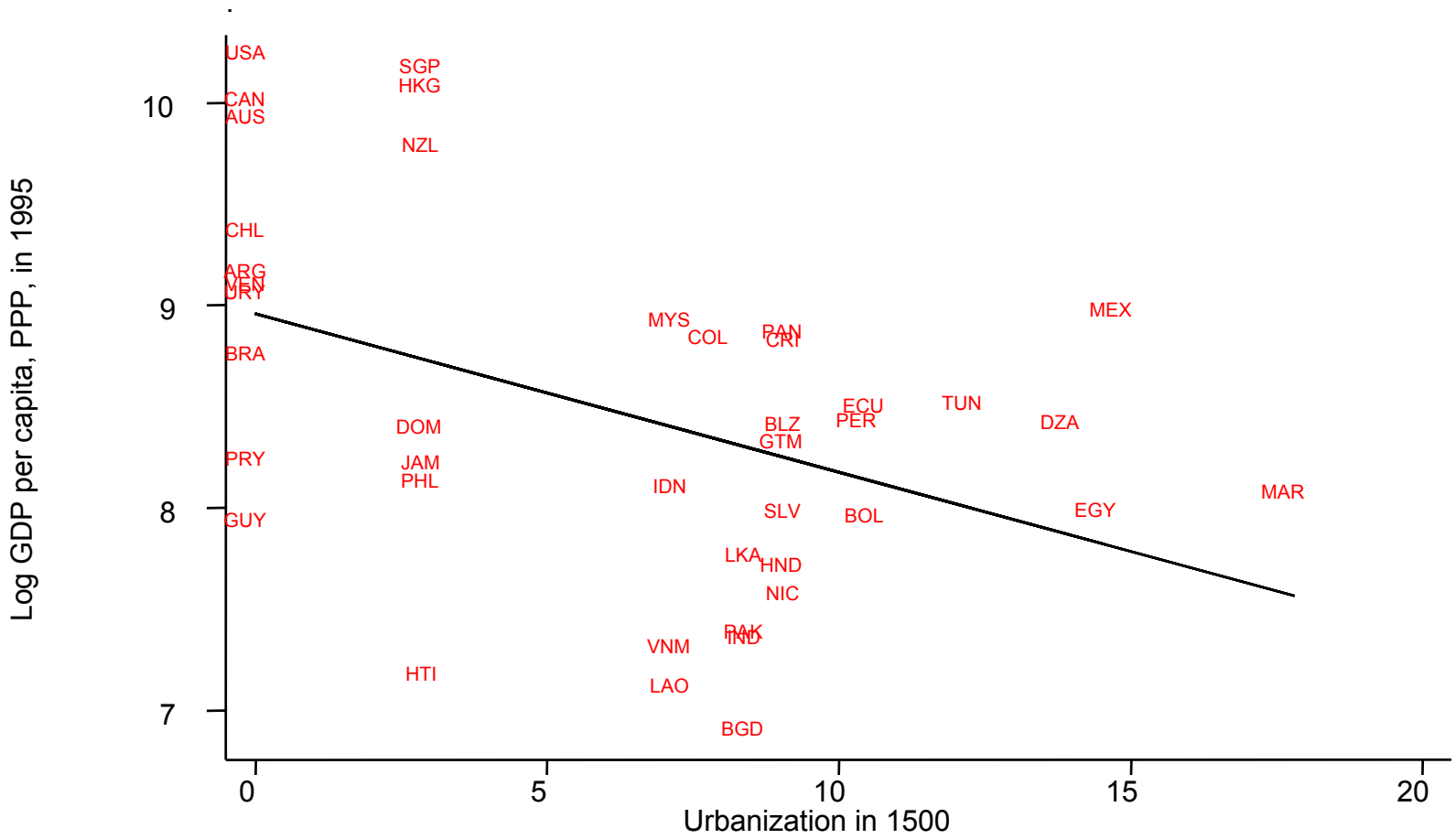


Figure 6
Log population density in 1500 and log GDP per capita in 1995, among former European colonies

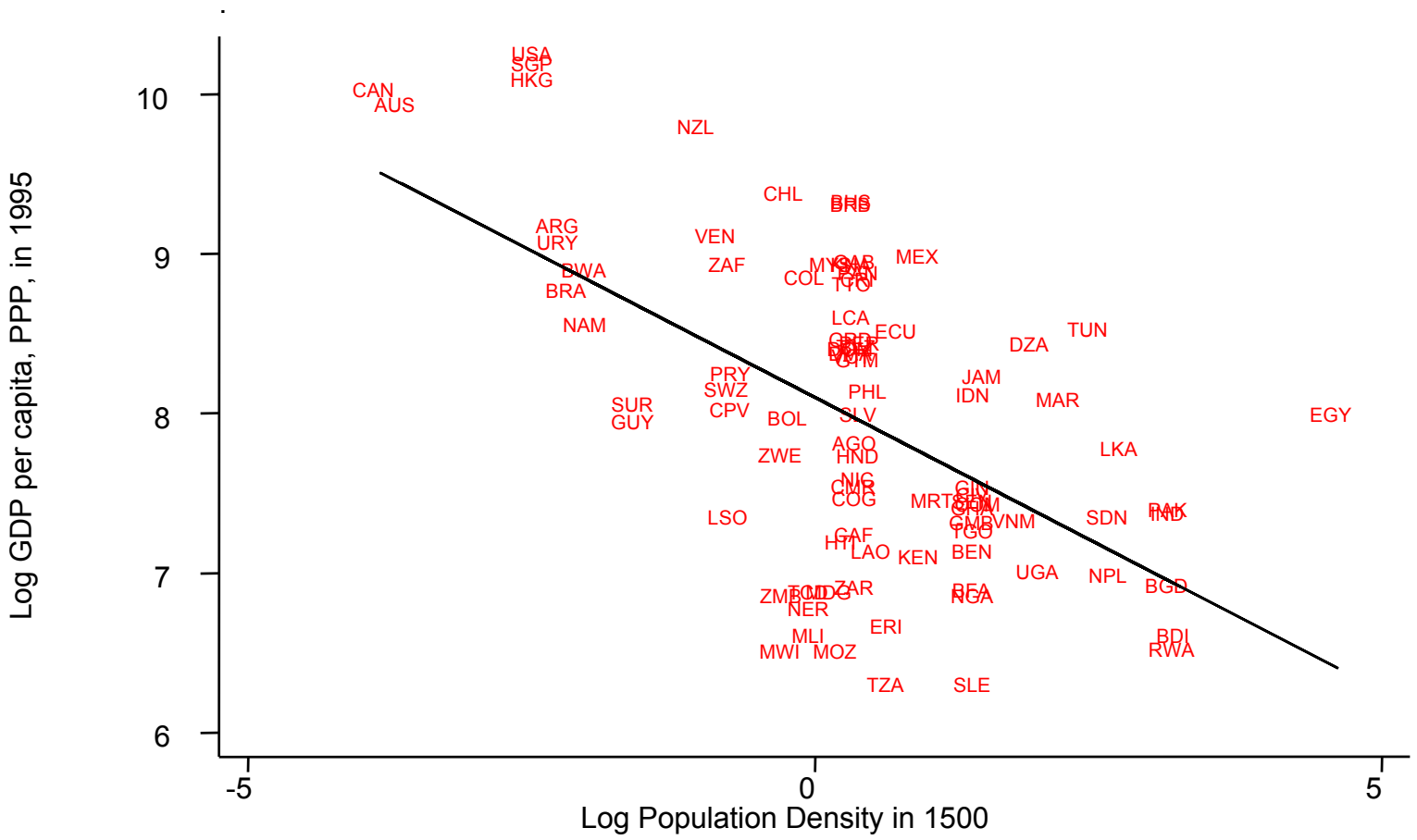


Figure 7
Urbanization in 1000 and 1500, among non-colonies

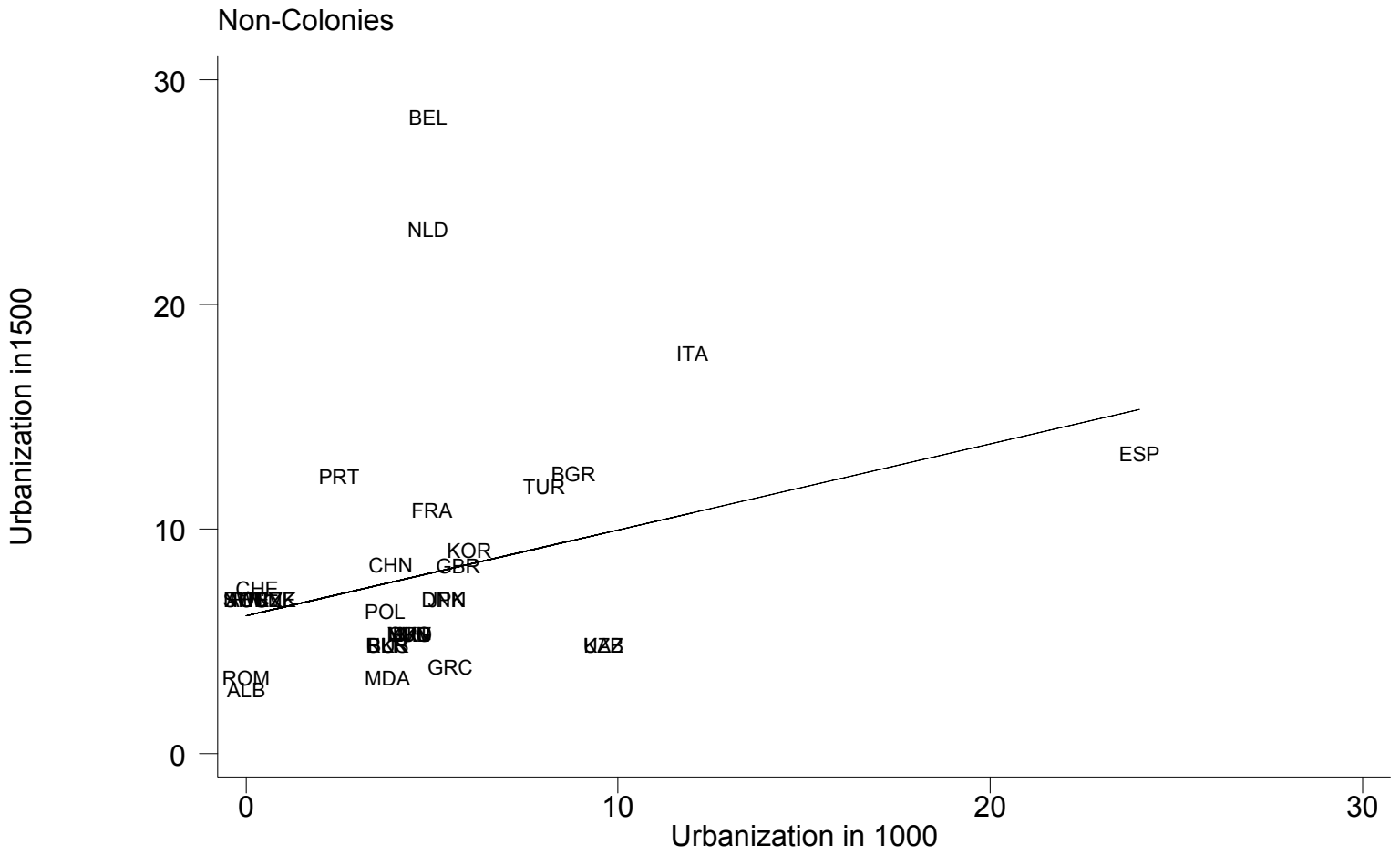


Figure 8
Urbanization in 1000 and 1500, among former European colonies

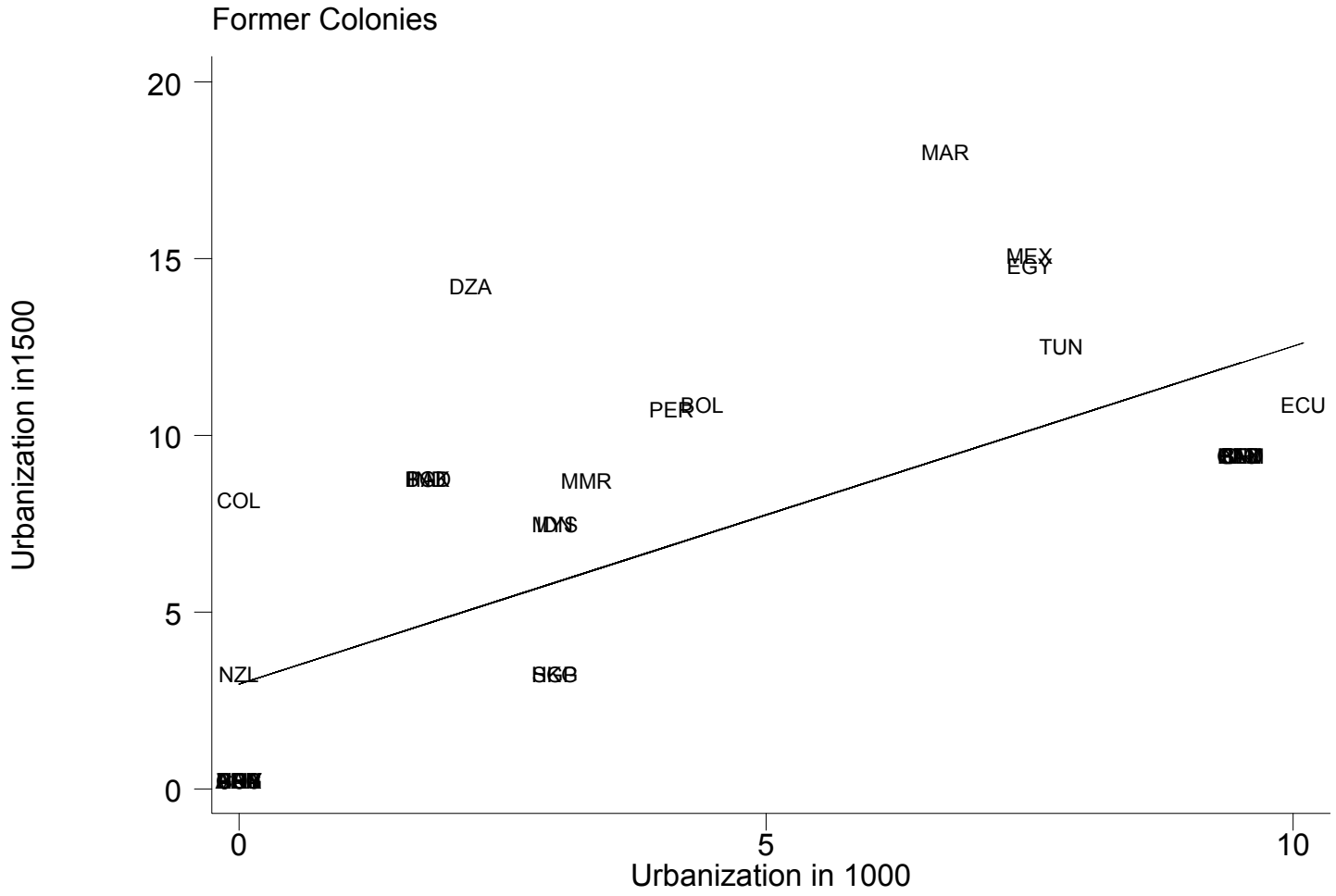


Figure 9
Urbanization in 1500 and log GDP per capita in 1995, among non-colonies

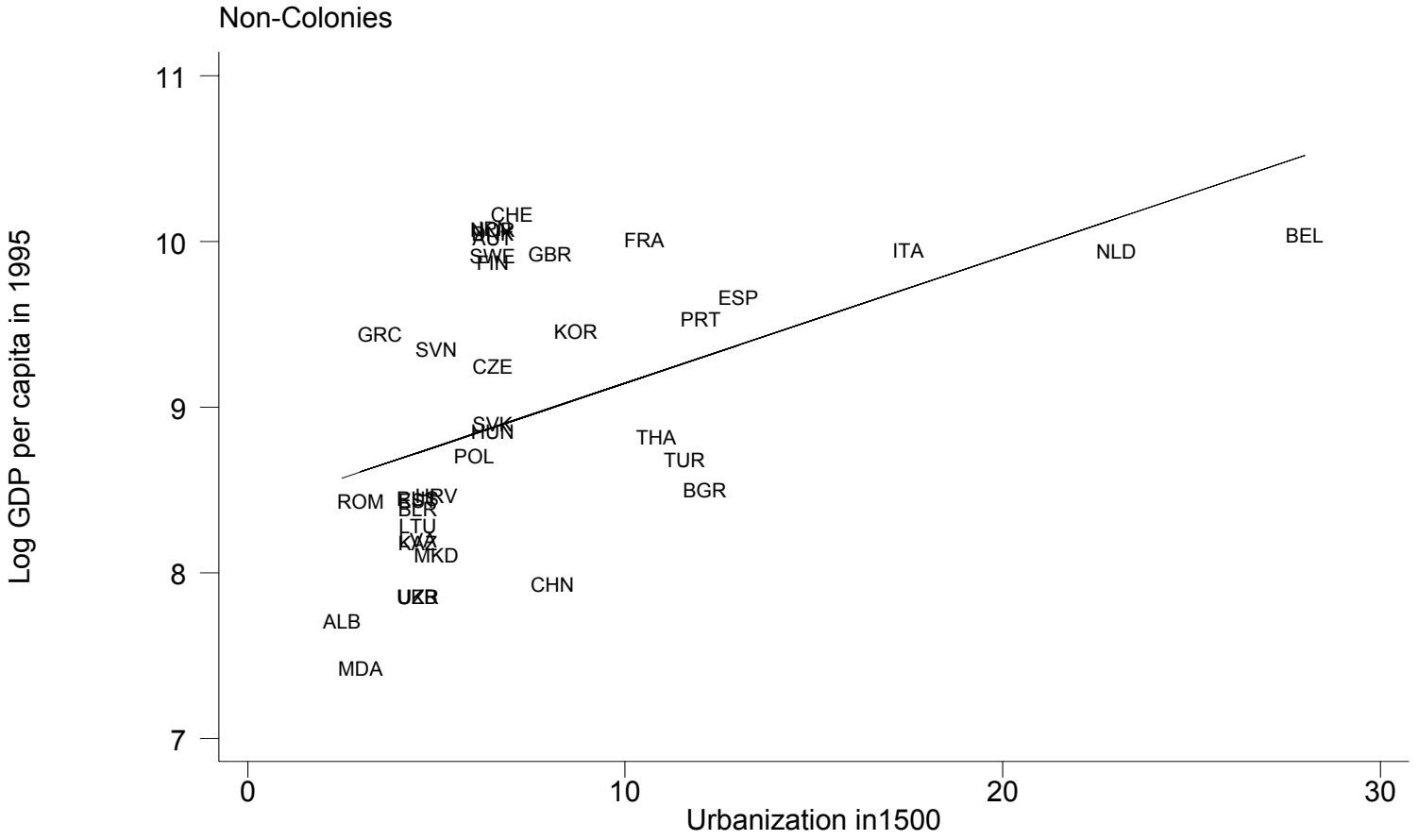


Figure 10
Evolution of urbanization among former European colonies

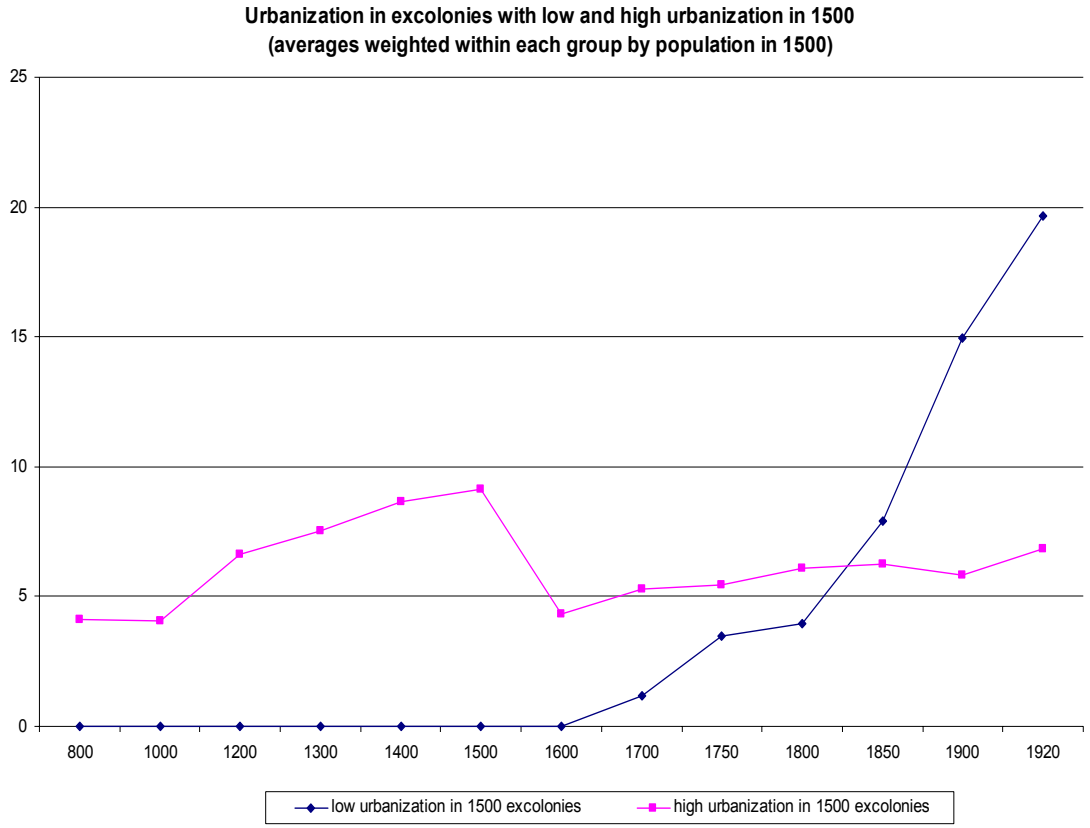


Figure 11
Evolution of industrial production per capita among former European colonies

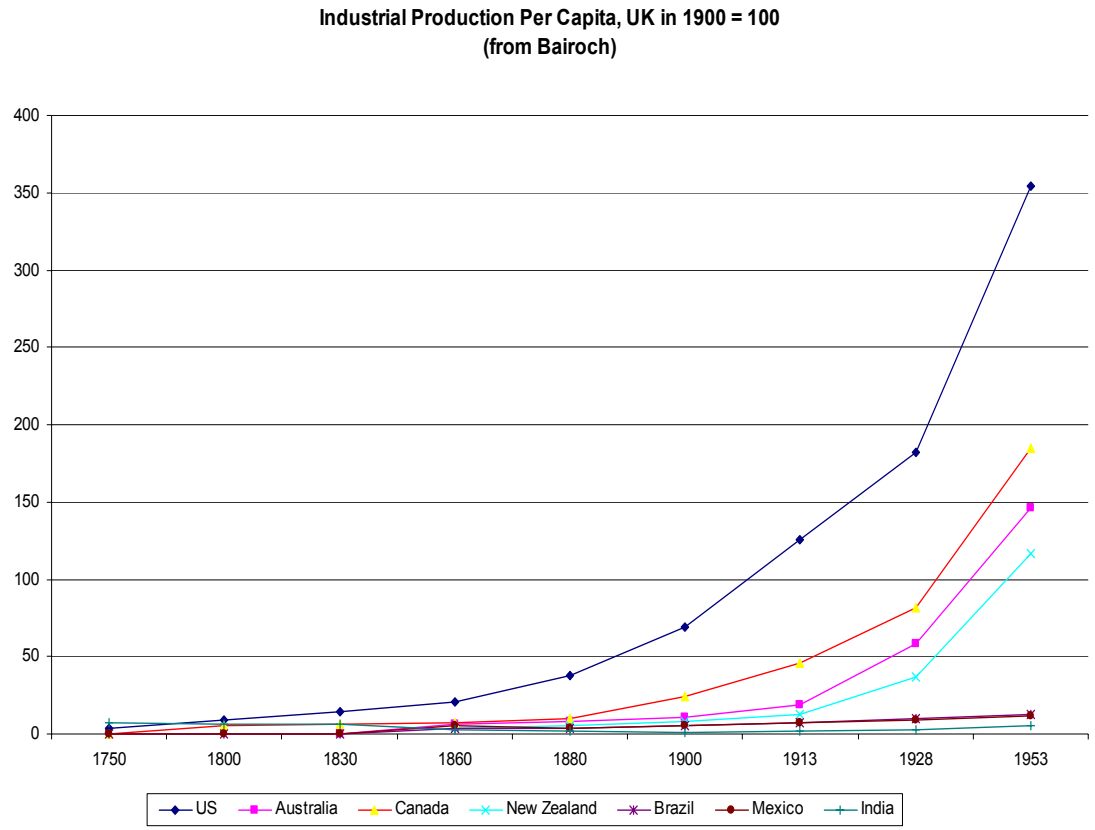


Figure 12
Urbanization in 1500 and average protection against risk of expropriation 1985-95

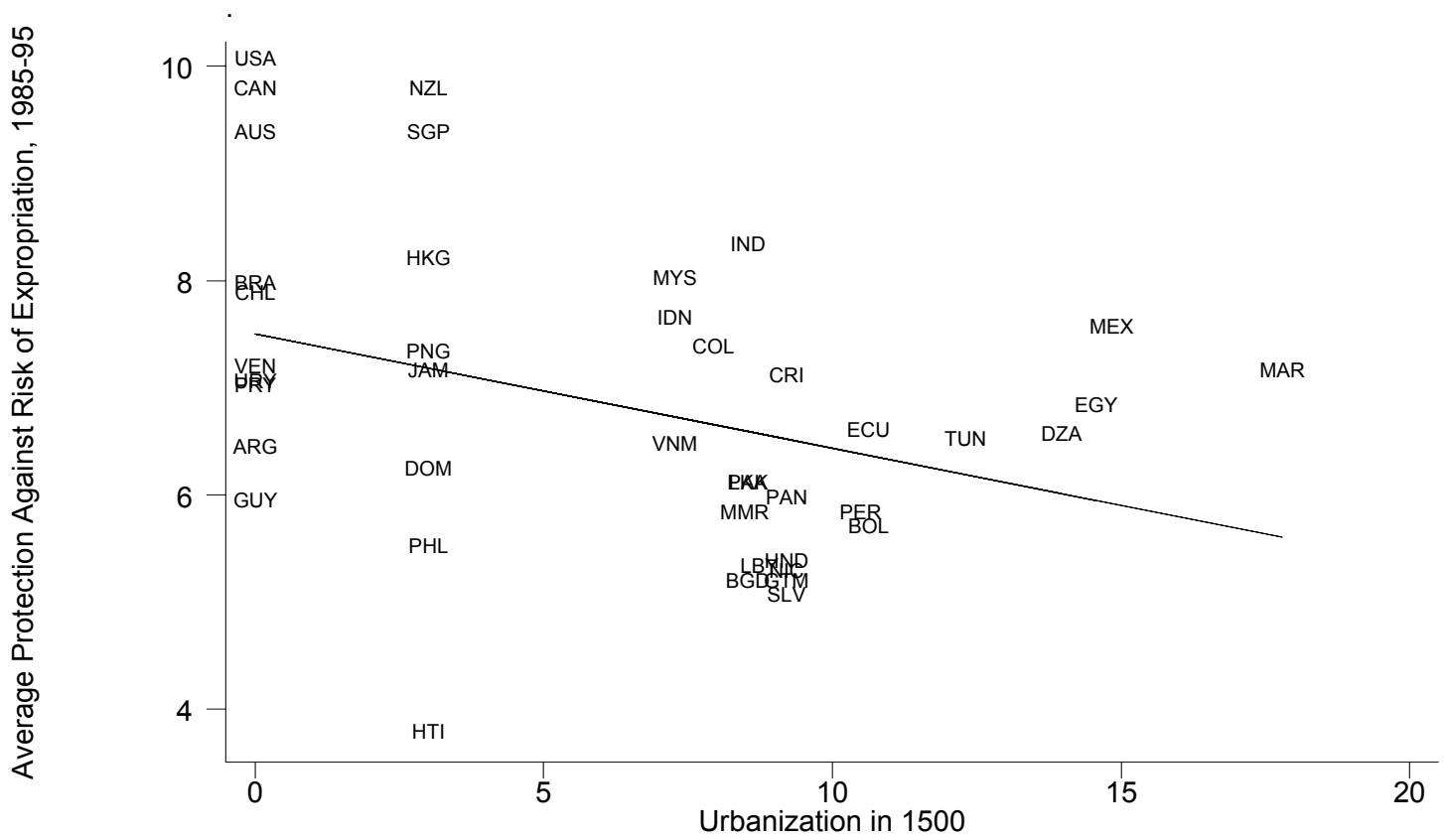


Figure 13
Log population density in 1500 and average protection against risk of expropriation 1985-95

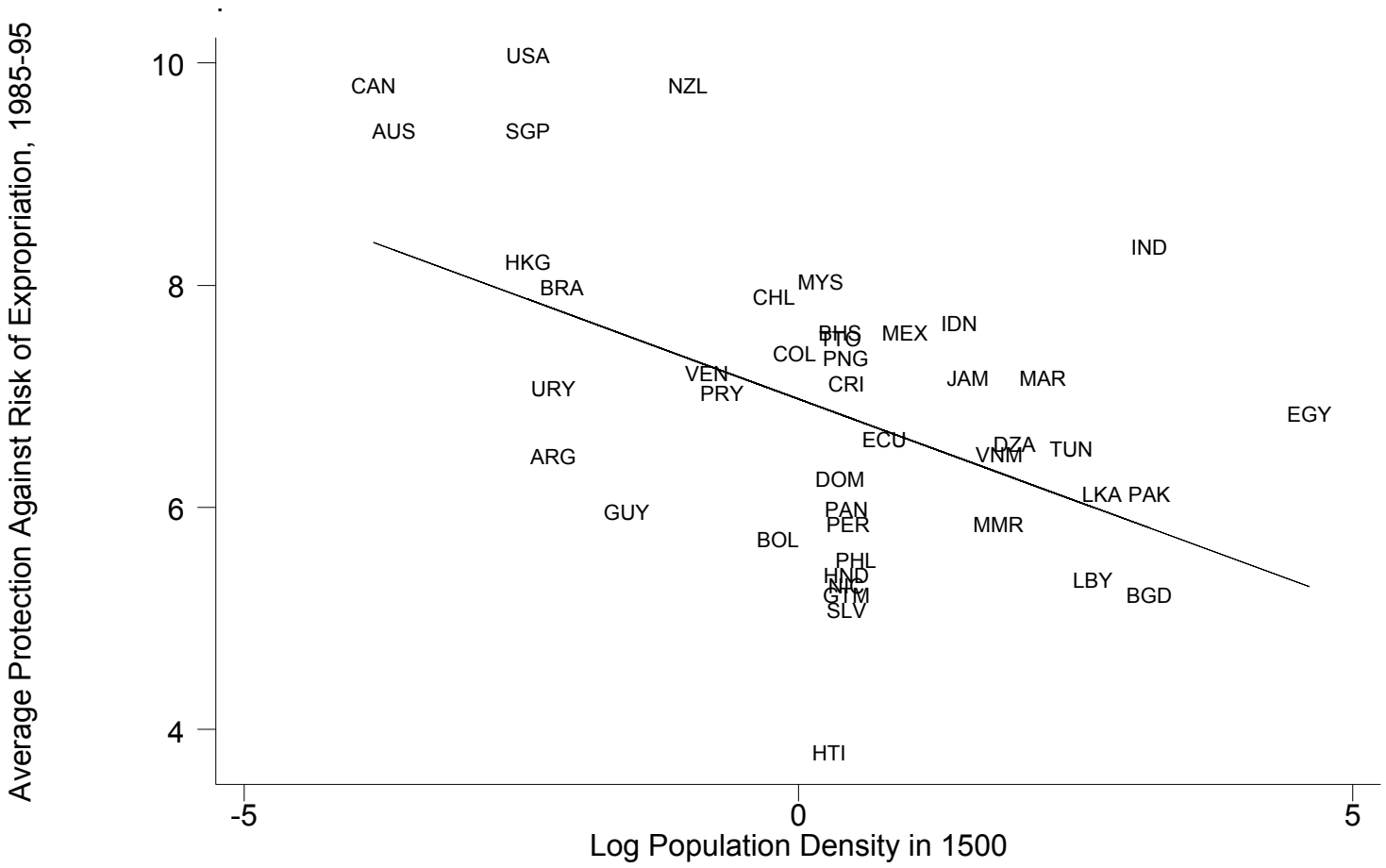


Figure 14
Log mortality of potential European settlers and average protection against risk of expropriation 1985-95

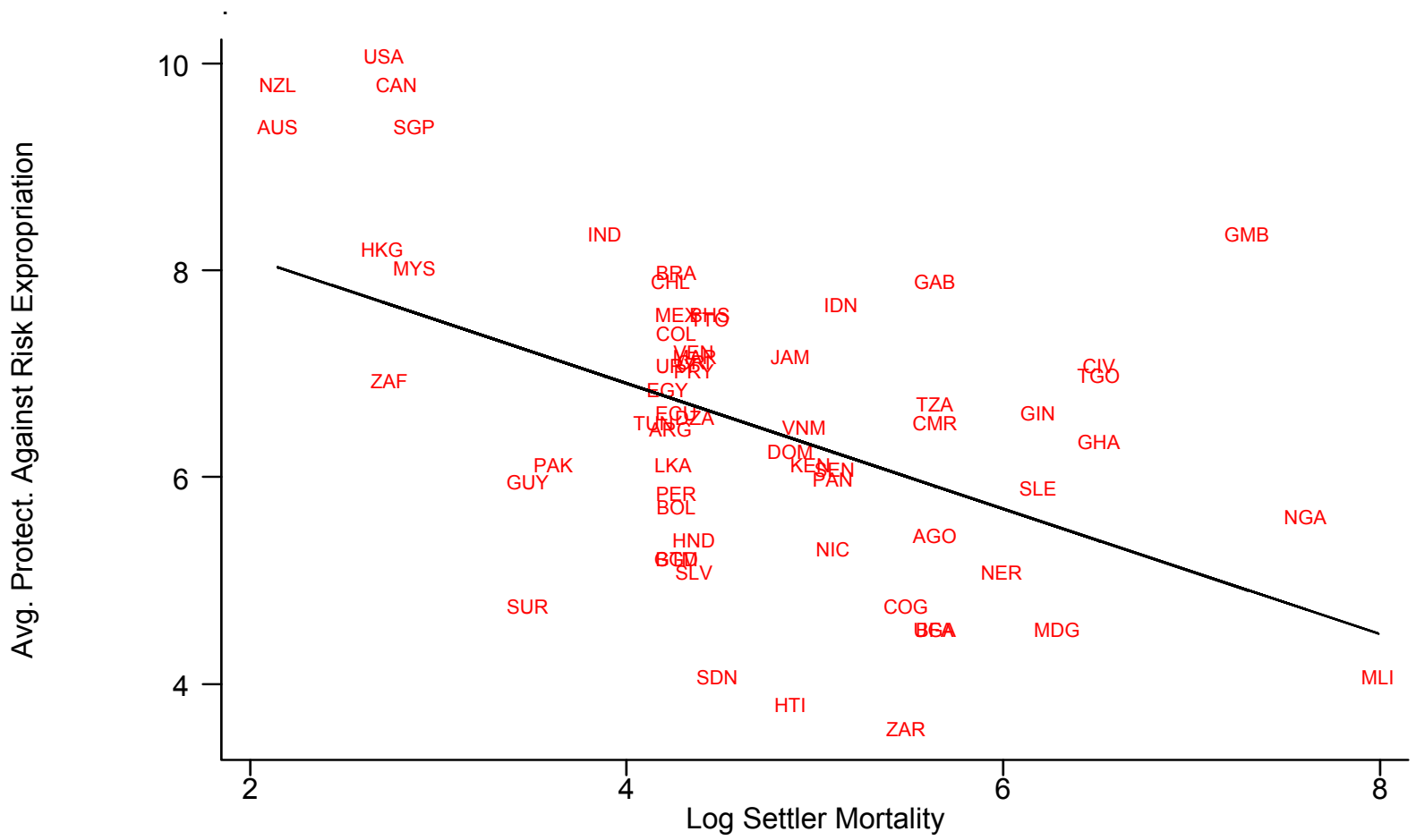


Figure 15
Log mortality of potential European settlers and log GDP per capita in 1995

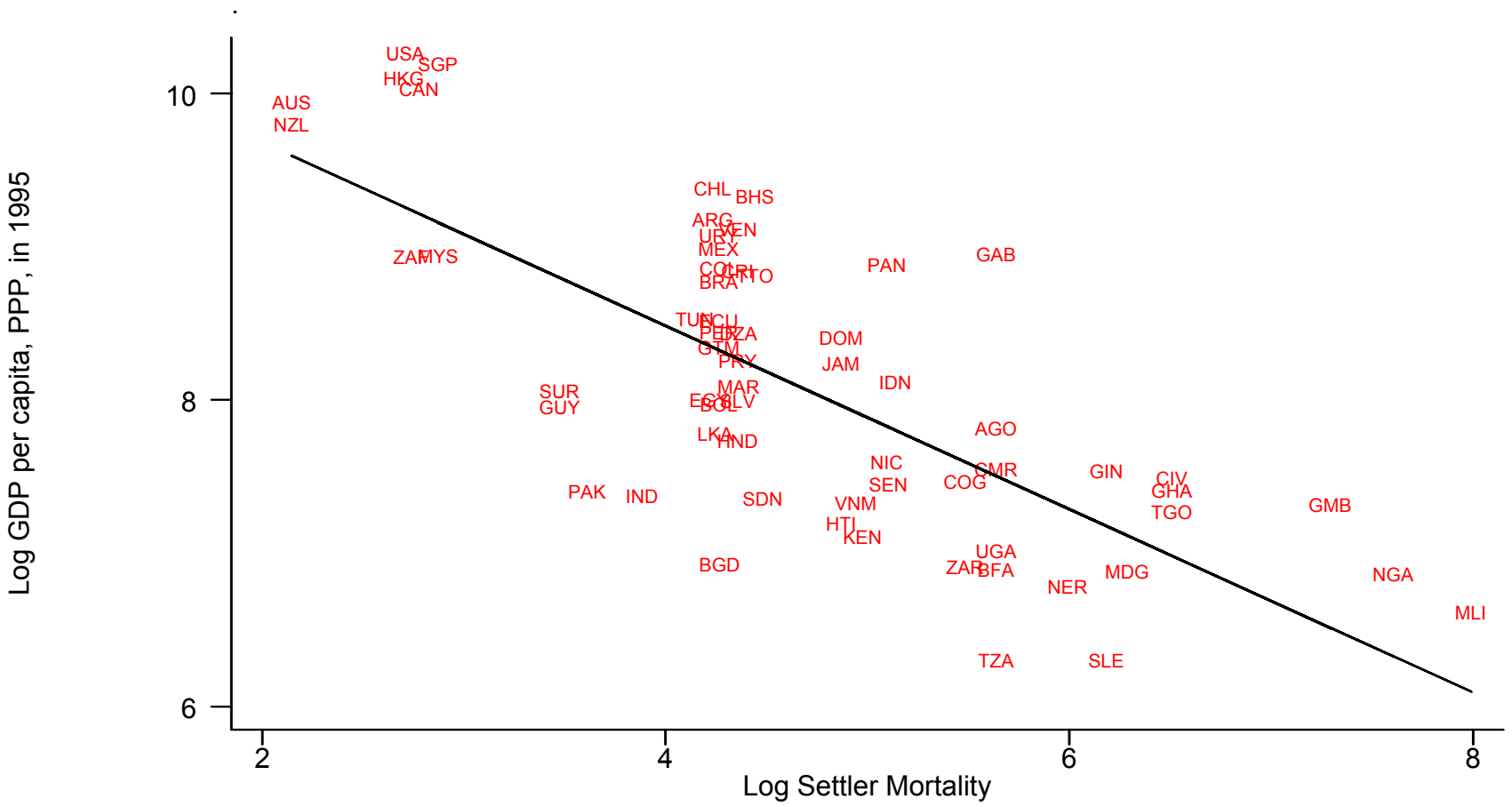


Figure 16
Log population density in 1500 and log GDP per capita in 1995, among former British colonies

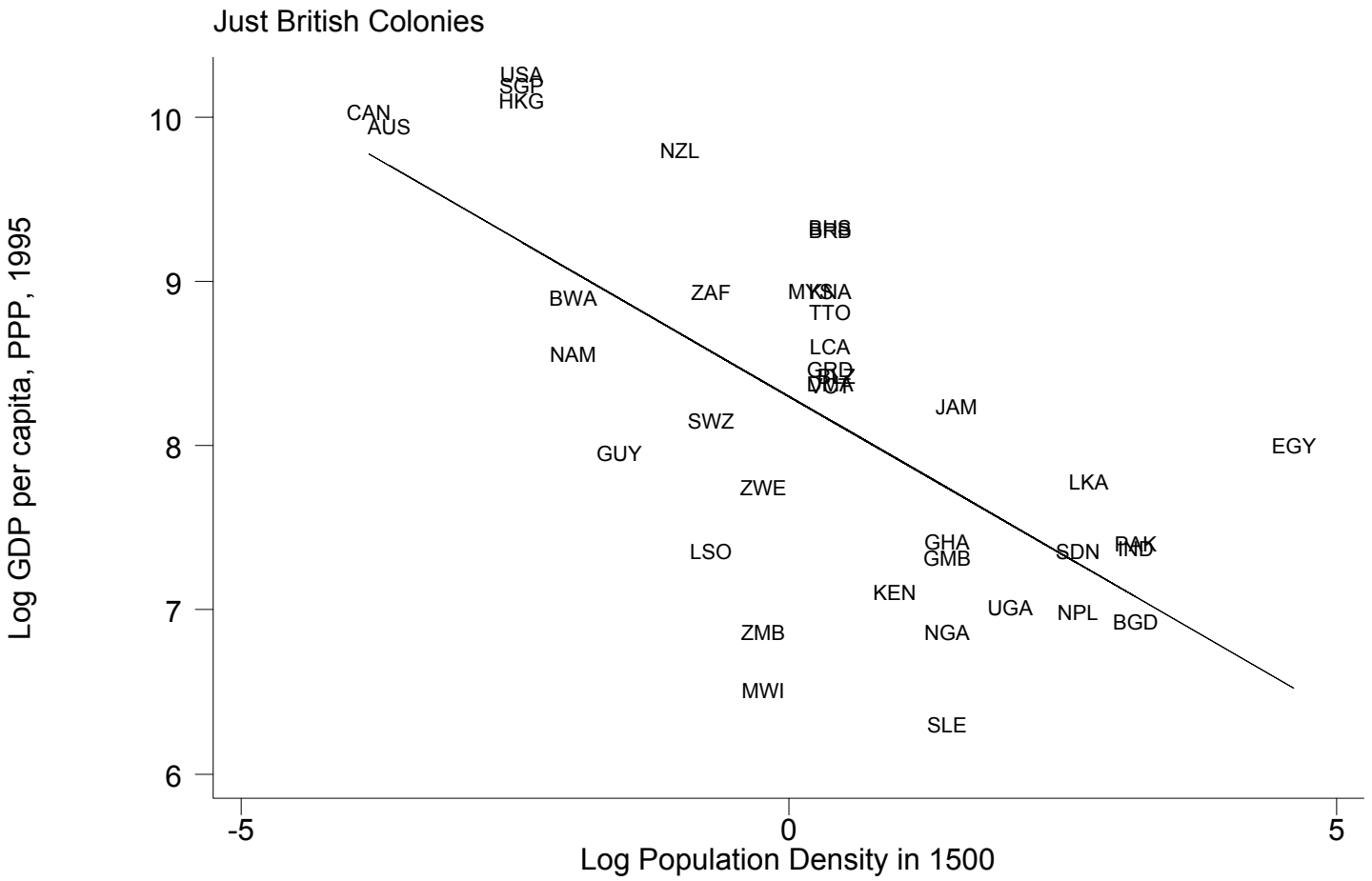
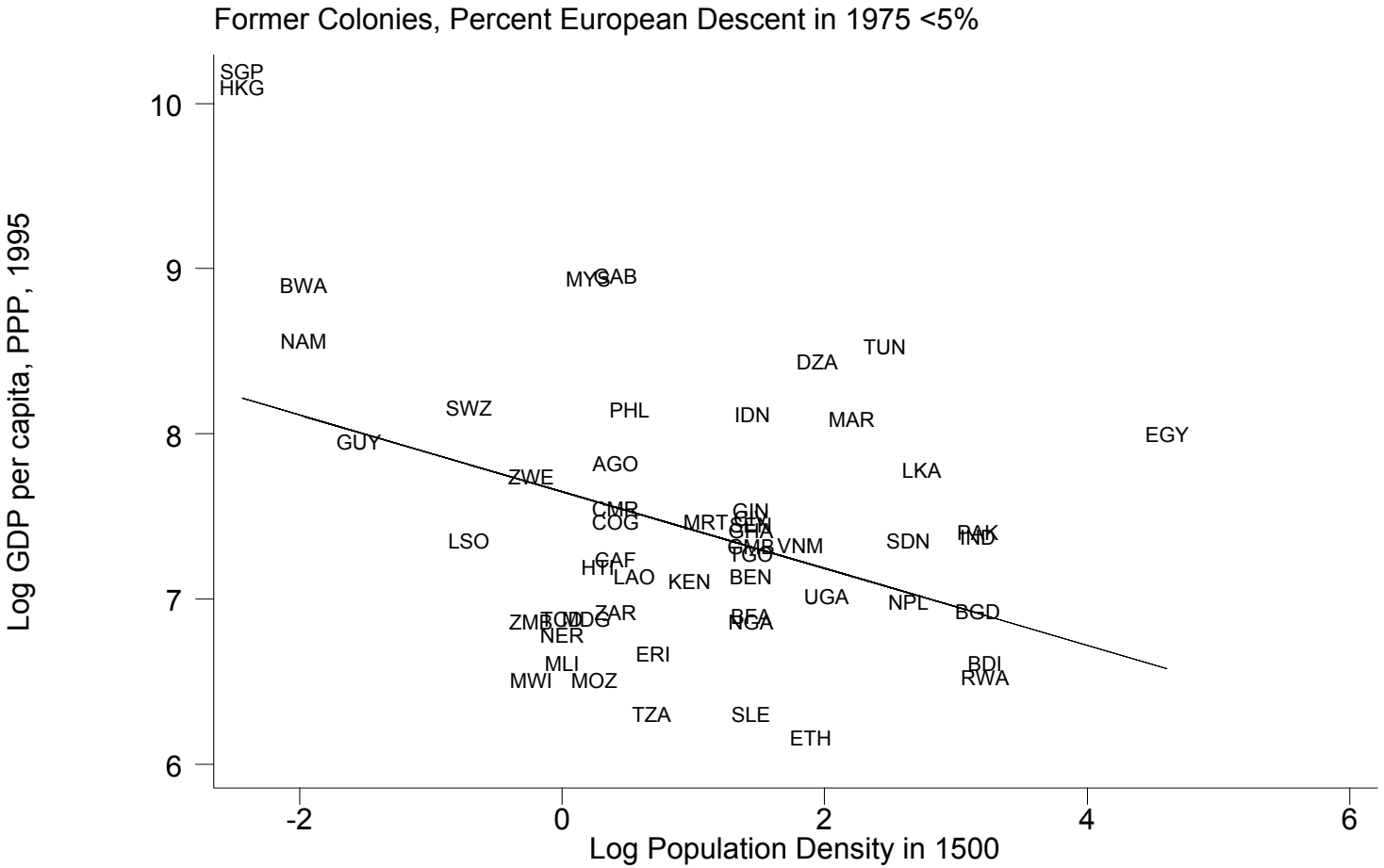


Figure 17
Log population density in 1500 and log GDP per capita in 1995, among former European colonies with current population less than 5% of European descent



Growth Theory Through the Lens of Development Economics

Abhijit V. Banerjee and Esther Duflo *

March 2004

1 Introduction: Neo-classical Growth Theory

The premise of neo-classical growth theory is that it is possible to do a reasonable job of explaining the broad patterns of economic change across countries, by looking at it through the lens of an aggregate production function. The aggregate production function relates the total output of an economy (a country, for example) to the aggregate amounts of labor, human capital and physical capital in the economy, and some simple measure of the level of technology in the economy as a whole. It is formally represented as $\mathcal{F}(A, \overline{K}_P \overline{K}_H, \overline{L})$ where $\overline{K}_P$ and $\overline{K}_H$ are the total amounts of physical and human capital invested, $\overline{L}$ is the total labor endowment of the economy and A is a technology parameter.

The aggregate production function is not meant to be something that physically exists. Rather, it is a convenient construct. Growth theorists, like everyone else, have in mind a world where production functions are associated with people. To see how they proceed, let us start with a model where everyone has the option of starting a firm, and when they do, they have access to an individual production function

$$Y = F(K_P, K_H, L, \theta), \tag{1}$$

where K_P and K_H are the amounts of physical and human capital invested in the firm and L is the amount of labor. θ is a productivity parameter which may vary over time, but at any point of time is a characteristic of the firm's owner. Assume that F is increasing in all its inputs. To make life simpler, assume that there is only one final good in this economy and physical capital is made from it. Also assume that the population of the economy is described by a distribution function $G_t(W, \theta)$, the joint distribution

*MIT, Department of Economics, 50 Memorial Drive, Cambridge, MA 02142. banerjee@mit.edu, eduflo@mit.edu. The authors are grateful to Michael Kremer, Rohini Pande, Chris Udry and Ivan Werning for helpful conversations and Seema Jayachandran for detailed comments. A part of this material was presented by the first author as the Kuznets Memorial Lecture, 2004, at Yale University. We are grateful for the many comments that we received from the audience. We gratefully acknowledge financial support from the Alfred P. Sloan Foundation.

of W and θ, where W is the wealth of a particular individual and θ is his productivity parameter. Let $\widetilde{G}(\theta)$ be the corresponding partial distribution on θ.

The lives of people, as often is the case in economic models, is rather dreary: In each period, each person, given his wealth, his θ and the prices of the inputs, decides whether to set up a firm, and if so how to invest in physical and human capital. At the end of the period, once he gets returns from the investment and possibly other incomes, he consumes and the period ends. The consumption decision is based on maximizing the following utility function:

$$\sum_{t=0}^{\infty} \delta^t U(C_t, \theta), 0 < \delta < 1. \tag{2}$$

1.1 The Aggregate Production Function

The key assumption behind the construction of the aggregate production function is that all factor markets are perfect, in the sense that individuals can buy or sell as much as they want at a given price. With perfect factor markets (and no risk) the market must allocate the available supply of inputs to maximize total output. Assuming that the distribution of productivities does not vary across countries, we can therefore define $\mathcal{F}(\overline{K}_P, \overline{K}_H, \overline{L})$ to be:

$$\begin{aligned} & \max_{\{K_P(\theta), K_H(\theta), L(\theta)\}} \{\int_\theta F(K_P(\theta), K_H(\theta), L(\theta), \theta) d\widetilde{G}(\theta)\} \\ \text{subject to} \int_\theta K_P(\theta) d\theta &= \overline{K}_P, \int_\theta K_H(\theta) d\theta = \overline{K}_H, and \int_\theta L(\theta) d\theta = \overline{L}. \end{aligned}$$

This is the aggregate production function. It is notable that the distribution of wealth does not enter anywhere in this calculation. This reflects the fact that with perfect factor markets, there is no necessary link between what someone owns and what gets used in the firm that he owns. The fact that $\widetilde{G}(\theta)$ does not enter as an argument of $\mathcal{F}(\overline{K}_P, \overline{K}_H, \overline{L})$ reflects our assumption that the distribution of productivities does not vary across countries.

It should be clear from the construction that there is no reason to expect a close relation between the "shape" of the individual production function and the shape of the aggregate function. Indeed it is well known that aggregation tends to convexify the production set: In other words, the aggregate production function may be concave even if the individual production functions are not. In this environment where there are a continuum of firms, the (weak) concavity of the aggregate production function is guaranteed as long as the average product of the inputs in the individual production functions is bounded in the sense that there is a λ such that $F(\lambda K_P, \lambda K_H, \lambda L, \theta) \leq \lambda \|(K_P, K_H, L, \theta)\|$ for all K_P, K_H, L and θ. It follows that the concavity of the individual functions is sufficient for the concavity of the aggregate but by no means necessary: The aggregate production would also be concave if the individual production functions

were S-shaped (convex to start out and then becoming concave). Alternately, the individual production function being bounded is enough to guarantee concavity of the aggregate production function. Moreover, the aggregate production function will typically be differentiable almost everywhere.

It is a corollary of this result that the easiest way to generate an aggregate production function with increasing returns is to base the increasing returns not on the shape of the individual production function, but rather on the possibility of externalities across firms. If there are sufficiently strong positive externalities between investment in one firm and investment in another, increasing the total capital stock in all of them together will increase aggregate output by more (in proportional terms) than the same increase in a single firm would raise the firm's output, which could easily make the aggregate production function convex. This is the reason why externalities have been intimately connected, in the growth literature, with the possibility of increasing returns.

The assumption of perfect factor markets is therefore at the heart of neo-classical growth theory. It buys us two key properties: The fact that the ownership of factors does not matter, i.e., that an aggregate production function exists; and that it is concave. The next sub-section shows how powerful these two assumptions can be.

1.2 The Logic of Convergence

Assume for simplicity that production only requires physical capital and labor and that the aggregate production function, $\mathcal{F}(\overline{K_p}, \overline{L})$ defined as above, exhibits constant returns and is concave, increasing, almost everywhere differentiable and eventually strictly concave, in the sense that $\mathcal{F}'' < \varepsilon < 0$, for any $\overline{K}_p > \widetilde{\overline{K_p}}$. As noted above, this does not require the individual production functions to have this shape, though it does impose some constraints on what the individual functions can be like. It does however require that the distribution of firm-level productivities is the same everywhere.

Under our assumption that capital markets are perfect, in the sense that people can borrow and lend as much as they want at the common going rate, r_t, the marginal returns to capital must be the same for everybody in the economy. This, combined with the fact that the preferences as represented by (2), has the immediate consequence that for everybody in the economy:

$$U'(C_t, \theta) = \delta r_t U''(C_{t+1}, \theta).$$

It follows that everybody's consumption in the economy must grow as long as $\delta r_t > 1$ and shrink if $\delta r_t < 1$. And since consumption must increase with wealth, it follows that everyone must be getting richer if and only if $\delta r_t > 1$, and consequently the aggregate wealth of the economy must be growing as long as $\delta r_t > 1$. In a closed economy, the total wealth must be equal to the total capital stock, and therefore the capital stock must also be increasing under the same conditions.

Credit market equilibrium, under perfect capital markets, implies that $\mathcal{F}'(\overline{K}_{Pt}, L) = r_t$. The fact that F is eventually strictly concave implies that as the aggregate capital stock grows, its marginal product must eventually start falling, at a rate bounded away from 0. This process can only stop when $\delta\mathcal{F}'(\overline{K}_{Pt}, L) = 1$. As long as the production function is the same everywhere, all countries must end up equally wealthy

The logic of convergence starts with the fact that in poor countries, capital is scarce, which combined with the concavity of the aggregate production function implies that the return on the capital stock should be high. Even with the same fraction of these higher returns being reinvested, the growth rate in the poorer countries would be higher. Moreover, the high returns should encourage a higher reinvestment rate, unless the income effect on consumption is strong enough to dominate. Together, they should make the poorer countries grow faster and catch up with the rich ones.

Yet poorer countries do not grow faster. According to Mankiw, Romer and Weil (1992), the correlation between the growth rate and the initial level of Gross Domestic Product is small, and if anything, positive (the coefficient of the log of the GDP in 1960 on growth rate between 1960 and 1992 is 0.0943). Somewhere along the way, the logic seems to have broken down.

Understanding the failure of convergence has been one of the key endeavors of the economics of growth. What we try to do in this chapter is to argue that the failure of this approach is intimately tied to the failure of the assumptions that underlie the construction of the aggregate production function and to suggest an alternative approach to growth theory that abandons the aggregate production.

We start by discussing, in section 2, the two implications of the neo-classical model that are at the root of the convergence result: Both rates of returns and investment rates should be higher in poor countries. We show that, in fact, neither rates of returns nor investment are, on average, much higher in poor countries. Moreover, contrary to what the aggregate production approach implies, there are large variations in rate of returns within countries, and large variation in the extent to which profitable investment opportunities are taken advantage of.

In section 3, we ask whether the puzzle (of low convergence) can be solved, while maintaining the aggregate production function, by theories that focus on reasons for technological backwardness in poor countries. We argue that this class of explanations is not consistent with the empirical evidence, which suggests that many firms in poor countries do use the latest technologies, while others in the same country use obsolete modes of production. In other words, what we need to explain is less the overall technological backwardness and more why some firms do not adopt profitable technologies that are available to them (though perhaps not affordable).

In section 4, we attempt to suggest some answers to the question of why firms and people in developing countries do not always avail themselves of the best opportunities afforded to them. We review

various possible sources of the inefficient use of resources: government failures, credit constraints, insurance failure, externalities, family dynamics, and behavioral issues. We argue that each of these market imperfections can explain why investment may not always take place where the rates of returns are the highest, and therefore why resources may be misallocated within countries. This misallocation, in turn, drives down returns and this may lower the overall investment rate. In section 5, we calibrate plausible magnitudes for the aggregate static impact of misallocation of capital within countries We show that, combined with individual production functions characterized by fixed costs, the misallocation of capital implied by the variation of the returns to capital observed within countries can explain the main aggregate puzzles: the low aggregate productivity of capital, and the low Total Factor Productivity in developing countries, relative to rich countries. Non-aggregative growth models thus seem to have the potential to explain why poor countries remain poor.

The last section provides an introduction to an alternative growth theory that does not require the existence of an aggregate production function, and therefore can accommodate the misallocation of resources. We then review the attempts to empirically test these models. We argue that the failure to take seriously the implications of non-aggregative models have led to results that are very hard to interpret. To end, we discuss an alternative empirical approach illustrated by some recent calibration exercises based on growth models that take the misallocation of resources seriously.

2 Rates of Return and Investment Rates in Poor Countries

In this section, we examine whether the two main implications of the neo-classical model are verified in the data: Are returns and investment rates higher in poor countries?

2.1 Are returns higher in poor countries?

2.1.1 Physical Capital

• Indirect Estimates

One way to look at this question is to look at the interest rates people are willing to pay. Unless people have absolutely no assets that they can currently sell, the marginal product of whatever they are doing with the marginal unit of capital should be no less than the interest rate: If this were not true, they could simply divert the last unit of capital toward whatever they are borrowing the money for and be better off.

There is a long line of papers that describe the workings of credit markets in poor countries (Banerjee (2003a) summarizes this evidence). The evidence suggests that a substantial fraction of borrowing takes

place at very high interest rates.

A first source of evidence is the "Summary Report on Informal Credit Markets in India" (Dasgupta (1989)), which reports results from a number of case studies that were commissioned by the Asian Development Bank and carried out under the aegis of the National Institute of Public Finance and Policy. For the rural sector, the data is based on surveys of six villages in Kerala and Tamil Nadu, carried out by the Centre for Development Studies. The average annual interest rate charged by professional moneylenders (who provide 45.6% of the credit) in these surveys is about 52%. For the urban sector, the data is based on various case surveys of specific classes of informal lenders, many of whom lend mostly to trade or industry. For finance corporations, they report that the minimum lending rate on loans of less than one year is 48%. For hire-purchase companies in Delhi, the lending rate was between 28% and 41%. For auto financiers in Namakkal, the lending rate was 40%. For handloom financiers in Bangalore and Karur, the lending rate varied between 44% and 68%.

Several other studies reach similar conclusions. A study by Timberg and Aiyar (1984) reports data on indigenous-style bankers in India, based on surveys they carried out: The rates for Shikarpuri financiers varied between 21% and 37% on loans to members of local Shikarpuri associations and between 21% and 120% on loans to non-members (25% of the loans were to non-members). Aleem (1990) reports data from a study of professional moneylenders that he carried out in a semi-urban setting in Pakistan in 1980-1981. The average interest rate charged by these lenders is 78.5%. Ghate (1992) reports on a number of case studies from all over Asia: The case study from Thailand found that interest rates were 5-7% per month in the north and northeast (5% per month is 80% per year and 7% per month is 125%). Murshid (1992) studies Dhaner Upore (cash for kind) loans in Bangladesh (you get some amount in rice now and repay some amount in rice later) and reports that the interest rate is 40% for a 3-5 month loan period. The Fafchamps (2000) study of informal trade credit in Kenya and Zimbabwe reports an average monthly interest rate of 2.5% (corresponding to an annualized rate of 34%) but also notes that this is the rate for the dominant trading group (Indians in Kenya, whites in Zimbabwe), while the blacks pay 5% per month in both places.

The fact that interest rates are so high could reflect the high risk of default. However, this does not appear to be the case, since several of studies mentioned above give the default rates that go with these high interest rates. The study by Dasgupta (1989) attempts to decompose the observed interest rates into their various components,[1] and finds that the default costs explain 7 per cent (not 7 percentage points!) of the total interest costs for auto financiers in Namakkal and handloom financiers in Bangalore and Karur, 4% for finance companies and 3% for hire-purchase companies. The same study reports that in four case studies of moneylenders in rural India they found default rates explained about 23% of the

[1] In the tradition of Bottomley (1963).

observed interest rate. Timberg and Aiyar (1984), whose study is also mentioned above, report that average default losses for the informal lenders they studied ranges between 0.5% and 1.5% of working funds. The study by Aleem gives default rates for each individual lender. The median default rate is between 1.5 and 2%, and the maximum is 10%.

Finally, it does not seem to be the case that these high rates are only paid by those who have absolutely no assets left. The "Summary Report on Informal Credit Markets in India" (Dasgupta (1989)) reports that several of the categories of lenders that have already been mentioned, such as handloom financiers and finance corporations, focus almost exclusively on financing trade and industry while Timberg and Aiyar (1984) report that for Shikarpuri bankers at least 75% of the money goes to finance trade and, to lesser extent, industry. In other words, they only lend to established firms. It is hard to imagine, though not impossible, that all the firms have literally no assets that they can sell. Ghate (1992) also concludes that the bulk of informal credit goes to finance trade and production, and Murshid (1992), also mentioned above, argues that most loans in his sample are production loans despite the fact that the interest rate is 40% for a 3-5 month loan period.

Udry (2003) obtains similar indirect estimates by restricting himself to a sector where loans are used for productive purpose, the market for spare taxi parts in Accra, Ghana. He collected 40 pairs of observations on price and expected life for a particular used car part sold by a particular dealer (e.g., alternator, steering rack, drive shaft). Solving for the discount rate which makes the expected discounted cost of two similar parts equal gives a lower bound to the returns to capital. He obtains an estimate of 77% for the median discount rate.

Together, these studies thus suggest that people are willing to pay high interest rates for loans used for productive purpose, which suggests that the rates of return to capital are indeed high in developing countries, at least for some people.

• Direct Estimates

Some studies have tried to come up with more direct estimates of the rates of returns to capital. The "standard" way to estimate returns to capital is to posit a production function (translog and Cobb-Douglas, generally) and to estimate its parameters using OLS regression, or instrumenting capital with its price. Using this methodology, Bigsten (2000) estimate returns to physical and human capital in five African countries. They estimate rates of returns ranging from 10% to 32%. McKenzie and Woodruff (2003) estimate parametric and non-parametric relationships between firm earnings and firm capital. Their estimates suggest huge returns to capital for these small firms: For firms with less than $200 invested, the rate of returns reaches 15% per *month*, well above the informal interest rates available in pawn shops or through micro-credit programs (on the order of 3% per month). Estimated rates of return

decline with investment, but remain high (7% to 10% for firms with investment between $200 and $500, 5% for firms with investment between $500 and $1,000).

Such studies present serious methodological issues, however. First, the investment levels are likely to be correlated with omitted variables. For example, in a world without credit constraints, investment will be positively correlated with the expected returns to investment, generating a positive "ability bias" (Olley and Pakes (1996)). McKenzie and Woodruff attempt to control for managerial ability by including the firm owner's wage in previous employment, but this may go only part of the way if individuals choose to enter self-employment precisely because their expected productivity in self-employment is much larger than their productivity in an employed job. Conversely, there could be a negative ability bias, if capital is allocated to firms in order to avoid their failure.

Banerjee and Duflo (2003a) take advantage of a change in the definition of the so-called "priority sector" in India to circumvent these difficulties. All banks in India are required to lend at least 40% of their net credit to the "priority sector", which includes small-scale industry, at an interest rate that is required to be no more than 4% above their prime lending rate. In January, 1998, the limit on total investment in plants and machinery for a firm to be eligible for inclusion in the small-scale industry category was raised from Rs. 6.5 million to Rs. 30 million. Banerjee and Duflo (2003a) first show that, after the reforms, newly eligible firms (those with investment between 6.5 million and 30 million) received on average larger increments in their working capital limit than smaller firms. They then show that the sales and profits increased faster for these firms during the same period. Putting these two facts together, they use the variation in the eligibility rule over time to construct instrumental variable estimates of the impact of working capital on sales and profits. After computing a non-subsidized cost of capital, they estimate that the returns to capital in these firms must be at least 94%.

There is also direct evidence of very high rates of returns on productive investment in agriculture. Goldstein and Udry (1999) estimate the rates of returns to the production of pineapple in Ghana. The rate of returns associated with switching from the traditional maize and Cassava intercrops to pineapple is estimated to be in excess of 1,200%! Few people grow pineapple, however, and this figure may hide some heterogeneity between those who have switched to pineapple and those who have not.

Evidence from experimental farms also suggests that, in Africa, the rate of returns to using chemical fertilizer (for maize) would also be high. However, this evidence may not be realistic, if the ideal conditions of an experimental farm cannot be reproduced on actual farms. Foster and Rosenzweig (1995) show, for example, that the returns to switching to high yielding varieties were actually low in the early years of the green revolution in India, and even negative for farmers without an education. This is despite the fact that these varieties had precisely been selected for having high yields, in proper conditions. But they required complementary inputs in the correct quantities and timing. If farmers were not able or did not

know how to supply those, the rates of returns were actually low.

Chemical fertilizer, however, is not a new technology, and the proper way to use it is well understood. To estimate the rates of returns to using fertilizer in actual farms in Kenya, Duflo and Robinson (2003), in collaboration with a small NGO, set up small scale randomized trials on people's farms: Each farmer in the trials delimited two small plots. On one randomly selected plot, a field officer from the NGO helped the farmer apply fertilizer. Other than that, the farmers continued to farm as usual. They find that the rates of returns from using a small amount of fertilizer varied from 169% to 500% depending on the year, although of returns decline fast with the quantity used on a plot of a given size.

The direct estimates thus tend to confirm the indirect estimates: While there are some settings where investment is not productive, there seems to be investment opportunities which yield substantial rates of returns.

• How high is the marginal product on average?

The fact that the marginal product in some firms is 50% or 100% or even more does not imply that the average of the marginal products across all firms is nearly as high. Of course, if capital always went to its best use, the notion of the average of the marginal products does not make sense. The presumption here is that there may be an equilibrium where the marginal products are not equalized across firms.

One way to get at the average of the marginal products is to look at the Incremental Capital Output Ratio (ICOR) for the country as a whole. The ICOR measures the increase in output predicted by a one unit increase in capital stock. It is calculated by extrapolating from the past experience of the country and assumes that the next unit of capital will be used exactly as efficiently (or inefficiently) as the last one. The inverse of the ICOR therefore gives an upper bound for the average marginal product for the economy—it is an upper bound because the calculation of the ICOR does not control for the effect of the increases in the other factors of production which also contributes to the increase in output.[2] For the late 1990s, the IMF estimates that the ICOR is over 4.5 for India and 3.7 for Uganda. The implied upper bound on the average marginal product is 22% for India and 27% in Uganda.

• Variations in the marginal products across firms.

To reconcile the high direct and indirect estimates of the marginal returns we just discussed and an average marginal product of 22% in India, it would have to be that there is substantial variation in the marginal product of capital within the country. Given that the inefficiency of the Indian public sector is legendary, this may just be explained by the investment in the public sector. However, since the ICOR

[2]The implicit assumption that the other factors of production are growing is probably reasonable for most developing countries, except perhaps in Africa.

is from the late 1990s, when there was little new investment (or even disinvestment) in the public sector, there must also be many firms in the private sector with marginal returns substantially below 22%. The micro evidence reported in Banerjee (2003b), which shows that there is very substantial variation in the interest rate within the same sub-economy, certainly goes in this direction. The Timberg and Aiyar (1984) study mentioned above, is one source of this evidence: It reports that the Shikarpuri lenders charged rates that were as low as 21% and as high as 120%, and some established traders on the Calcutta and Bombay commodity markets could raise funds for as little as 9%. The study by Aleem (1990), also mentioned above, reports that the standard deviation of the interest rate was 38.14%. Given that the average lending rate was 78.5%, this tells us that an interest rate of 2% and an interest rate of 150% were both within two standard deviations of the mean. Unfortunately, we cannot quite assume from this that there are some borrowers whose marginal product is 9% or less: The interest rate may not be the marginal product if the borrowers who have access to these rates are credit constrained. Nevertheless, given that these are typically very established traders, this is less likely than it would be otherwise.

Ideally we would settle this issue on the basis of direct evidence on the misallocation of capital, by providing direct evidence on variations in rates of return across groups of firms. Unfortunately such evidence is not easy to come by, since it is difficult to consistently measure the marginal product of capital. However, there is some rather suggestive evidence from the knitted garment industry in the Southern Indian town of Tirupur (Banerjee and Munshi (2004); Banerjee and Munshi (2003)). Two groups of people operate in Tirupur: the Gounders, who issue from a small, wealthy, agricultural community from the area around Tirupur, who have moved into the ready-made garment industry because there was not much investment opportunity in agriculture. Outsiders from various regions and communities started joining the city in the 1990s. The Gounders have, unsurprisingly, much stronger ties in the local community, and thus better access to local finance, but may be expected to have less natural abilities for garment manufacturing than the outsiders, who came to Tirupur precisely because of its reputation as a center for garment export. The Gounders own about twice as much capital as the outsiders on average. They maintain a higher capital-output ratio than the outsiders at all levels of experience, though the gap narrows over time. The data also suggest that they make less good use of their capital than the outsiders: While the outsiders start with lower production and exports than the Gounders, their experience profile is much steeper, and they eventually overtake the Gounders at high levels of experience, even though they have lower capital stock throughout. This data therefore suggests that capital does not flow where the rates of return are highest: The outsiders are clearly more able than the Gounders, but they nevertheless invest less.[3]

[3]This is not because capital and talent happen to be substitutes. In this data, as it is generally assumed, capital and ability appear to be complements.

To summarize, the evidence on returns to physical capital in developing countries suggests that there are instances with high rates of return, while the average of the marginal rates of return across firms does not appear to be that high. This suggests a coexistence of very high and very low rates of return in the same economy.

2.1.2 Human Capital

• Education

The standard source of data on the rate of return to education is Psacharopoulos (1973; 1985; 1994; 2002)) who compiles average Mincerian returns to education (the coefficient of years of schooling in a regression of log(wages) on years of schooling) as well as what he call "full returns" to education by level of schooling. Compared to Mincerian returns, full returns take into account the variation in the cost of schooling according to year of schooling: The opportunity cost of attending primary school is low, because 6 to 12 year old children do not earn the same wage as adults; and the direct costs of education increase with the level of schooling.

On the basis of this data, Psacharopoulos argues that returns to education are substantial, and that they are larger in poor countries than in rich countries. We re-examine the claim that returns to education are larger in poor countries, using data on traditional Mincerian returns, which have the advantage of being directly comparable. We start with the latest compilation of rates of returns, available in Psacharopoulos (2002) and on the World Bank web site. We update it as much as possible, using studies that seem to have been overlooked by Psacharopoulos, or that have appeared since then (the updated data set and the references are presented in the appendix).[4] We flag the observations that Bennell (1996) rated as being of "poor" or "very poor" quality. We complete this updated database by adding data on years of schooling for the year of the study when it was not reported by Psacharopoulos.

Using the preferred data, the Mincerian rates of returns seem to vary little across countries: The mean rate of returns is 8.96, with a standard deviation of 2.2. The maximum rate of returns to education (Pakistan) is 15.4%, and the minimum is 2.7% (Italy). Averaging within continents, the average returns are highest in Latin America (11.05) and lowest in the Europe and the U.S. (7%), with Africa and Asia in the middle.

If we run an OLS regression of the rates of returns to education on the average educational attainment (number of years of education), using the preferred data (updated database without the low quality data), the coefficient is -0.26, and is significant at 10% level (table 1, column 3). The returns to education

[4]The bulk of the update is for African countries, where Bennell (1996) had systematically investigated the Psacharoupoulos data, and found that many of the underlying studies were unreliable.

predicted from this regression ranges from 6.91 for the country with the lowest education level to 10.09 for the country with the highest education level. This is a small range (smaller than the variation in the estimates of the returns to education of a single country, or even in different specifications in a single paper!): There is therefore no *prima facie* evidence that returns to education are much higher when education is lower, although the relationship is indeed negative. Columns 1 and 2 in the same table show that the data construction matters: When the countries with "poor" quality are included, the coefficient of years of education increases to -0.45. When only the 38 countries in the latest Psacharopoulos update are included (most countries are dropped because the database does not report years of education, even for countries where it is clearly available–Austria for example), the coefficient more than doubles, to -0.71. On the whole, this strong negative number does appear to be an artifact of data quality.

In column (4), we directly regress the Mincerian returns to education on GDP, and we find a small and significant negative relationship. However, this is counteracted by the fact that teacher salary grows less fast than GDP, and the cost of education is thus not proportional to GDP: In column (5) we regress the log of the teacher salary on the log of GDP per capita.[5] The coefficient is significantly less than one, suggesting that teachers are relatively more expensive in poor countries. This is to some extent attenuated by the fact that class sizes are larger in poor countries (which tends to make education cheaper). We then compute the returns to educating a child for one year as the ratio of the lifetime benefit of one year of education (assuming a life span of 30 years, a discount rate of 5%, a share of wage in GDP of 60%, and no growth), to the direct cost of education (assuming that teacher salary is 85% of the cost of education). In column (6), we regress this ratio on GDP: There is no relationship between this measure of returns and GDP.[6] If we factor in indirect costs (as a fraction of GDP) (in column 7), the relationship becomes slightly more negative, but still insignificant. On balance, the returns to one more year of education are therefore no higher in poor countries.

• Health

Education is not the only dimension of human capital. In developing countries, investment in nutrition and health has been hypothesized to have potentially high returns at moderate levels of investment. The report of the Commission for Macroeconomics and Health (on Macroeconomics and Health (2001)), for example, estimated returns to investing in health to be on the order of 500%, mostly on the basis of cross-country growth regressions. Several excellent recent surveys by John Strauss and Duncan Thomas (Strauss and Thomas (1995); Strauss and Thomas (1998)), Thomas (2001) and Thomas and Frankenberg (2002) summarize the existing literature on the impact of different measures of health on fitness and

[5]The teacher salary data is obtained from the "Occupational Wages Around the World" database (Freeman and Oostendorp (2001)).

[6]Note that by assuming that the lifespan is the same in poor and rich countries, we are biasing upwards the returns in poor countries.

productivity, and lead to a much more nuanced conclusion.

There is substantial experimental evidence that supplementation in iron and vitamin A increases productivity at relatively low cost. Unfortunately, not all studies report explicit rates of returns calculations. The few numbers that are available suggest that some basic health intervention can have high of returns: Basta and Scrimshaw (1979) studies an iron supplementation experiment conducted among rubber tree tappers in Indonesia. Baseline health measures indicated that 45% of the study population was anemic. The intervention combined an iron supplement and an incentive (given to both treatment and control groups) to take the pill on time. Work productivity in the treatment group increased by 20% (or $132 per year), at a cost per worker-year of $0.50. Even taking into account the cost of the incentive ($11 per year), the intervention suggests extremely high rates of returns. Duncan Thomas and Al (2003) obtain lower, but still high, estimates in a larger experiment, also conducted in Indonesia: They found that iron supplementation experiments in Indonesia reduced anemia, increased the probably of participating in the labor market, and increased earnings of self-employed workers. They estimate that, for self-employed males, the benefits of iron supplementation amount to $40 per year, at a cost of $6 per year.[7] The cost benefit analysis of a de-worming program (Basta and Scrimshaw (1979)) in Kenya reports estimates of a similar order of magnitude: Taking into account externalities (due to the contagious nature of worms), the program led to an average increase in school participation of 0.14 years. Using a reasonable figure for the returns to a year of education, this additional schooling will lead to a benefit of $30 over the life of the child, at a cost of $0.49 per child per year. Not all interventions have the same rates of return however: A study of Chinese cotton mill workers (Li and Hautvast (1994)) led to a significant increase in fitness, but no corresponding increase in productivity. Likewise, the intervention analyzed by Duncan Thomas and Al (2003) had no effect on earnings or labor force participation of women.

In summary, while there is not much debate on the impact of fighting anemia (through iron supplementation or de-worming) on work capacity, there is more heterogeneity amongst estimates of economic rates of return of these interventions. The heterogeneity is even larger when we consider other forms of health interventions, reviewed, for example, in Strauss and Thomas (1995), or when one compares various human capital interventions. As in the case of physical capital, there are instances of high returns, and substantial heterogeneity in returns.

[7]This number takes into account the fact that only 20% of the Indonesian population is iron deficient: The private returns of iron supplementation for someone who knew they were iron deficient–which they can find out using a simple finger prick–would be $200).

2.1.3 Taking Stock: Returns on Capital

The marginal product of physical and human capital in developing countries seems very high in some instances, but not necessarily uniformly. The average of the marginal products of physical capital in India may well be less than 22%, though even reasonably large firms often have marginal products of 60%, or even 100%.

The question is whether we should think of 22% as a high number or a low number. One way to think about it is that it is only 2.5 times the 9% or so that a marginal dollar earns in the U.S. (the average stock market real return), but is variation by a factor of 2.5 as much we might ever expect?

A more structured way to answer this question is to follow Lucas (1990), and to ask whether, in the neo-classical model, the marginal product of capital is high enough in India to be compatible with the observed difference in output-per-worker. According to the Penn World Tables (Heston and Aten (2002)), in 1990, output-per-worker in India at Purchasing Power Parity was 1/11th of what it was in the U.S. To obtain a productivity gap per effective use of labor, we need to adjust this ratio by the differences in education between the two countries. Based on the work of Krueger (1967), Lucas (1990) argues that "one American worker is equal to five Indian workers" in terms of human capital. In our case, since we are comparing productivity in 1990, and Krueger's estimates of human capital are from the late 1960s, we presumably adjust the correction factor. Between 1965 and 1990, years of schooling among those 25 years or older went from 1.90 years to 3.68 years in India and from 9.25 years to 12 years in the United States, i.e., from approximately 20% of the U.S. level, which fits with the 5:1 gap in productivity that Krueger suggested, to about 30%.[8]

To show what this implies, Lucas starts with the assumption that *net* output is produced using a production function $Y = AL^{1-\alpha}K^{\alpha}$, where K is investment and L is the number of worker.[9]

From this, it follows that ouput per worker is $y = Ak^{\alpha}$, where k is investment per worker in equipment. Assuming that firms can borrow as much as they want at the rate r, profit maximization requires that $\alpha Ak^{\alpha-1} = r$, from which it follows that

$$\frac{y_U}{y_I} = \left(\frac{r_I}{r_U}\right)^{\frac{\alpha}{1-\alpha}} \left(\frac{A_U}{A_I}\right)^{\frac{1}{1-\alpha}}. \tag{3}$$

If we assume that the only difference between the TFP levels in the two countries is due to the

[8]These numbers are based on Barro and Lee (2000). Another angle from which this can be looked at is that health improved also during the period: Over a slightly different period, (1970-75 to 1995-2000), according to the Human Development Report (United Nations Development Program (2001)), life expectancy at birth went from 50.3 to 62.3 years in India and from 71.5 years to 76.5 years in the U.S., reducing the gap between India and the U.S. by about 40%.

[9]Lucas actually computes the ratio of output per effective unit of labor, which, with our parameters, is equal to $11 * \frac{3}{10} \approx 3$. Reassuringly, this is also the ratio that Lucas started with, albeit based on the average numbers for the 1965-1990 period rather than the 1990 numbers.

productivity per worker, the fact that Indian workers are only 30% as productive as the US workers and the share of capital is assumed to be 40% implies that:

$$\frac{A_U}{A_I} = (0.3)^{0.6} \approx 2. \tag{4}$$

With these parameters, the 11-fold difference between y_U and y_I would imply that $r_I = (3)^{5/3} r_U \approx 5r_U$. r is naturally thought of as the marginal product of capital. Given the difference in output-per-worker even after adjusting for the difference in workers' productivity, the marginal product of capital in India should therefore be 5 times, rather than 2.5 times, what it is in the US. In other words, if we take 9% for the marginal product of capital in the U.S., this would imply a 45% rate for India, rather than the 22% we observe in the data.

Lucas saw this as an obvious reason to reject the assumption that the TFP levels in the two countries are the same on the grounds that if the rates were indeed that different, capital would flow from the U.S. to India. Hence, Lucas argued, the rate of returns cannot possibly be that high in India. This seems to be rather a leap of faith: As we have already seen, there are indeed many investment opportunities in India that yield 45% or more, and capital flows from within or without have not yet eliminated them.

On the other hand, it is clear that if the average marginal rate in India is really 22% or less (which is our best guess about it), then Lucas was right in insisting that the actual rates of returns are much lower than what we would expect if the model were correct. Likewise, while there are some types of investment in human capital in developing countries that yield high rates of return, the average investment in human capital does not appear to be particularly lucrative.

The data thus contradicts the predictions of the aggregate production approach from two different directions: First, rates of return are very far from being equalized. Second, the average of these returns is lower in poor countries than an aggregate production function would predict, under the assumption of equal TFP. We now turn to other implications of the neo-classical model: Investment rates should be higher in poor countries and, within countries, capital should be flowing to investments that yield the highest return.

2.2 Investment Rates in Poor Countries

2.2.1 Is investment higher in poor countries?

Prima facie, it does not seem to be the case that investment rates are higher in poor countries. On the contrary, there is a robust positive correlation between investment rates in physical capital and income per capita, when both are expressed in terms of purchasing power parity. In fact, Levine and Renelt (1992) and Sala-I-Martin (1997) identified investment per capita as the only robust correlate of income.

For example, Hsieh and Klenow (2003) estimate that in 1985, the correlation between PPP investment rate and PPP income per capita for the 115 countries present in the Penn World Tables was 0.60. The coefficients they estimate suggest that an increase in one log point in income per capita is associated with about a 5 percentage point higher PPP investment rate (the mean investment rate is 14.5%). The same positive correlation obtains with investment in plant and machinery. The relationship between investment rate and income per capita is much less strong when both of them are expressed in nominal terms rather than in PPP terms (Eaton and Kortum (2001); Restuccia and Urrutia (2001) and Hsieh and Klenow (2003)). The coefficient drops by a third when all investments are considered, and becomes insignificant when the measure of investment includes only plant and machinery. According to Hsieh and Klenow (2003), the fact that poor countries have a lower investment-to-GDP ratio, when expressed in PPP, is explained by the low relative price of consumption, relative to investment: While there is no correlation between investment prices and GDP, there is a strong positive correlation between consumption prices and GDP. It is not clear, however, that knowing this helps us explain why there is not *more* investment in poor countries. First, because the high rates that we found in some firms in developing countries and the lower, but still high, rates that we found on average are there despite the high price of capital goods. Moreover, even if we measure everything in nominal terms, there is no strong negative correlation between investment and GDP.

There are, of course, examples of poor countries with large investment to GDP ratios. Young (1995) shows that a substantial fraction of the rapid growth of the East-Asian economies in the post-WWII period can be accounted for by rapid factor accumulation (including increase in the size of the labor force, factor reallocation, and high investment rates). In particular, according to the national accounts, between 1960 and 1985, the capital stock in Singapore, Korea, and Taiwan grew at more than 12% a year (in Hong Kong, it grew only at 7.7% a year). Between 1966 and 1999, the capital-output ratio has increased at an average rate of 3.4% a year in Korea, and 2.8% in Singapore. In Singapore, for example, the constant investment-to-GDP ratio increased from 10% in 1960 to 47% in 1984. In Singapore, Korea, and Taiwan, this increase in the stock of capital alone is responsible for about 1% out of the average yearly 3.4% to 4% of the "naive" Solow residual. Based on these results, Alwyn Young (Young (1995)) concluded that the East-Asian economies are perfect examples of transitional dynamics in the neo-classical model. However, in subsequent research, Hsieh (1999) questioned the validity of the national account data for investment for Singapore. He observes that if the capital-to-GDP ratio had grown at that speed, one would have observed a commensurate reduction in the rental price of capital. In practice, there was indeed a steady fall in the rental price of capital (both the interest rates and the relative price of capital fell) in Korea, Taiwan and Hong Kong. The drop is particularly large in Korea, where the national account statistics also suggest a large increase in the capital stock. However, in Singapore, there is no

evidence that the rental rate declined over the period. If any thing, it seems to have *increased*.

As for investment in physical capital, there is no *prima facie* evidence that poor countries invest more in education. The data is poor and extremely partial, since it is difficult to estimate private expenditure on education. What we can measure easily, government expenditure on education as a fraction of GDP, however, is not higher in poor countries, though there is significant variation across countries. In 1996, according to the country level data disseminated by the World Bank "edstat" department, government investment on education was 4.8% in Africa, 4% in Asia, 4.1% in Latin America, 4.8% in North America and 5.6% in Europe. The correlation between the log of government expenditure on education as a fraction of GDP and GDP-per-capita is strong (in current prices): The coefficient of the log of GDP was 0.18 in 1990, and 0.08 in 1996, larger than the comparable estimate for rate of investment in physical capital.

As we noted earlier, the fact that teachers are relatively more expensive in developing countries may imply that true returns to education may be much lower than the Mincerian returns. Can this explain why there is not greater investment in education in poor countries? Within the neo-classical model, the answer is no: Banerjee (2003b) shows that in the neo-classical world the same forces that raise the relative price of teachers in poor countries (or in countries with low education levels) also raise the wages paid to educated people, and on net the rate of return has to be higher rather than lower. And, in any case, it is not true that public investment in education is higher when returns are higher: We found no correlation between government expenditure on education as a fraction of GDP and rate of returns to education (the coefficient of the rates of return to education on government expenditure in education in 1996 is -0.008, with a standard error of 0.013).

In summary, while there are isolated cases of high investment rates in relatively poor countries (Taiwan and Korea), this by no means seems to be a general phenomenon. We have already suggested one reason why this might be the case–it does not look like returns are especially high. It may also be that investment is not particularly responsive with respect to returns. This is the issue we turn to next.

2.2.2 Does investment respond to rates of return?

There is little doubt that people do take up many investment opportunities with high potential returns. Investment flowed into Bangalore when it became a hub for the software industry in India. When, in the 1990s, Tirupur, a smallish town in South India, became known in the U.S. as a good place to contract large orders of knitted garments, the industry in the city grew at more than 50% per year, due to substantial investments of both the local community (diversifying out of agriculture) and outsiders attracted to Tirupur (Banerjee and Munshi (2004)). Or, to take a last example from India, new hybrid seeds and fertilizers spread rapidly during the "green revolution", leading to very rapid yield growth

(yields were multiplied by 3 in Karnataka and 2.5 in Punjab (Foster and Rosenzweig (1996))).

However, there are many instances where investments options with very high rates of returns do not seem to be taken advantage of. For example, Goldstein and Udry (1999) find that, despite the high rates of returns to growing pineapple compared to other crops, only 18% of the land is used for pineapple farming. Similarly, Duflo and Robinson (2003) find that only less than 15% of maize farmers in the area where they conducted field trials on the profitability of fertilizer report having used fertilizer in the previous season, despite estimated rates of return in excess of 100%.

From a more macro perspective, Bils and Klenow (2000) argue that the observed high correlation between educational attainment and subsequent growth observed in cross-sectional data (one year of additional schooling attainment is associated with 0.30 percent faster annual growth over the period 1960-1990) must be due, at least in part, to the fact that higher expected growth rates increase the returns to schooling, and therefore the demand for schooling. As we noted earlier, the correlation between education and subsequent growth (found in many studies, e.g., Barro (1991), Benhabib and Spiegel (1994), and Barro and Sala-I-Martin (1995)) appears to be too high to be entirely explained by the causal effect of transitional differences in human capital growth rates on growth rates. Bils and Klenow (2000) calibrate a simple neo-classical growth model, which requires that the impact of schooling on individual productivity has to be consistent with the average coefficient obtained from Mincer regressions. Their calibration suggest that the high level of education in 1960 can only explain up to a third of the correlation between education and growth. Moreover, as we discussed above, this correlation cannot be explained by high human capital externalities. They therefore calibrate an alternative model, where they construct the optimal schooling predicted by a country's expected economic growth. The calibration, once again, requires that the impact of education on human capital be consistent with the micro-estimates of the Mincerian returns, so that there remains a large fraction of the correlation between education and growth to explain. Higher expected growth induces more schooling by lowering the effective discount rate. They assume that a country's expected growth is a weighted average of its real *ex post* growth and the growth of the rest of the world. They estimate that, starting at 6.2 years of schooling, a 1 percent increase in growth induces 1.4 to 2.5 more years of schooling, depending on the values chosen for the parameters that are imposed. A 1 percentage point higher Mincerian return to schooling increases education by 1.1 to 1.9 years.

The aggregate data is thus consistent with a strong response of schooling to growth. However, it is also consistent with the presence of an omitted variable explaining both education and growth: In fact, Bils and Klenow acknowledge that their estimates suggests an elasticity of schooling demand to returns to schooling that is higher than what is implied by existing micro-studies (reviewed by Freeman (1986)). This problem cannot really be adequately addressed in the macroeconomic data, since there it is difficult

to find a plausible instrument for growth, and the impact of expected growth on schooling must essentially be estimated as a residual impact (what remains to be explained from the correlation between growth and schooling after a plausible estimate for the impact of education on growth has been removed).

Foster and Rosenzweig, in a series of papers, use the green revolution in India as a source of partly exogenous increase in rate of returns to human capital to estimate the impact of expected growth and increases in returns to education on schooling and, more generally, investment in human capital. Foster and Rosenzweig (1996) find that returns to education increased faster in regions where the green revolution induced faster technological change: Their estimates imply that in 1971, before the start of the green revolution, the profits in households where the head had completed primary education were 11% higher than the profits in households were he had not. By 1982, the profits were 46% higher for districts where the growth rate was one standard deviation above average. They then turn to estimating whether educational choices were also sensitive to the higher yield growth. After instrumenting for yield growth, they find that the impact of technological change on education is indeed substantial: In areas with recent growth in yields of one standard deviation above the mean, the enrollment rates of children from farm households are an additional 16 percentage points (53%) higher, compared to average-growth areas. Foster and Rosenzweig (2000) find that technological growth also affected the provision of schools, benefiting landless households. However, on balance, technological growth seems to lead to lower educational investment by landless households, perhaps because returns to education increase less for them (since they are engaged in more menial tasks) and because the fact that the withdrawal of children of landed households from the labor market increases children's wages, and thus the opportunity cost of school attendance.

Foster and Rosenzweig (1999) consider another measure of investment in children's human capital, namely child survival. They argue that technological growth in the village increases the returns from investing in boys' health, while technological growth outside the village, but in the potential "marriage market", increases the returns to investing in girls (because better educated and healthier women will fetch a higher prices in regions with higher technological progress). Their results indeed suggest that the gap in boys/girls mortality rates increases with technological change in the village, but decreases with technological change in the labor market.

Other evidence that girls' survival is affected by the expected returns to having girls include Rosenzweig and Schultz (1982), who show that the boys/girls mortality gap is negatively correlated to women's wage, and Qian (2003), who uses the liberalization of tea prices in China as a natural experiment in female productivity. She shows that, in regions suitable to tea production, the ratio of boys to girls diminished considerably after tea production and tea prices were liberalized: She interprets this as evidence that prospects for higher productivity for girls (women are particularly suited to tea picking) encourage parents to invest in their girls.

While these facts taken together do suggest that individuals do respond to returns when making human capital investment decisions, there are possible alternative explanations for these facts. The results from Rosenzweig and Schultz (1982) and Qian (2003) cannot easily be distinguished from a women's bargaining power effect: If mothers tend to prefer girls, and their bargaining power increases as a result of the increase of their productivity, then the outcomes will improve for girls, even if households' decisions do not respond to returns. The results in Foster and Rosenzweig (1996, 2000) could in part be attributed to wealth effects (expected growth makes the households richer, and if education has any consumption value, one would expect growth to respond to it), although Foster and Rosenzweig (1996) estimate the wealth effect directly, and argue that it is not important. But it remains possible that the instrumented expected increase in yield captures real increases in expected wealth better than any other measure (they show that land prices do adjust to the future expected yield increases, for example). Moreover, there is also direct evidence that investment in human capital does not always respond to returns: Munshi and Rosenzweig (2004) show that the rapid increase in the returns to English education in India in the 1990s (the returns increased from 15% to 24% in 10 years for boys, and 0% to 27% for girls) led to a convergence in the choice of English as a medium of instruction between the low and high castes amongst girls, but not amongst boys: Boys from the lower castes seem so far not to have taken full advantage of the new opportunities offered by English medium education.

Another angle for approaching this question is the sensitivity of human capital investment to the direct or indirect costs of these investments. Several recent studies do suggest that the elasticity of school participation with respect to user fees is high: Kremer and Namunyu (2003) conducted a randomized trial in rural Kenya in which an NGO provided uniforms, textbooks, and classroom construction to seven schools randomly selected from a pool of 14 schools. Dropouts fell considerably in the schools that received the program, relative to the other schools (after five years, pupils initially enrolled in the treatment schools had completed 15% more schooling than those enrolled in the comparison schools). They argue that the financial benefits of the free uniforms were the main reason for this increase in participation. Several programs go beyond reducing the school fees to actually pay for attendance. The PROGRESA program in Mexico provided grants to poor families, conditional on continued school participation and participation in health care. The program was initially launched as a randomized experiment, with 506 communities randomly assigned to either the treatment or control group. Schultz (2001) finds a 3.4% increase in enrollment in all children. The largest increase was in the transition between primary and secondary school, especially for girls. Gertler and Boyce (2002) report a similar effect on health. In this case as well, it is difficult to distinguish the pure price effect from the income effect.[10] School meals, which is another way to pay children to attend school, have been shown to be associated with increased school

[10]Moreover, there could be a bargaining power effect, since the grants were distributed through women.

participation in several observational studies (Jacoby (2002); Long (1991); Powell and Elston (1983); Powell and Grantham-McGregor (1998) and Dreze and Kingdon (1999)) and one experimental trial conducted among pre-school children in Kenya (Vermeersch (2002)). The available evidence, therefore, points toward a robust elasticity of schooling decisions with respect to the cost of schooling.

While this could be indicative of households being extremely sensitive to net returns, the magnitude of these effects are hard to reconcile with this explanation. For example, using an estimate of 7% Mincerian returns per year of education, Miguel and Kremer (2004) estimate that the benefit of one year of primary schooling is in excess of $200 over the lifetime of a child. Yet, the provision of a uniform valued at $6 induced an average increase of 0.5 years in the time a child spent in school (time spent in schools increased from 4.8 years in the comparison schools on average, to 5.3 years in the treatment schools). To be consistent with a model where the only reason where the provision of uniforms increase school attendance is the increase in the rate of returns that it leads to, these numbers would mean that a large fraction of children (or their parents) were exactly indifferent between attending school or not, before the uniform is provided.

While this is certainly possible, other evidence suggests that human capital investment does not always respond to rates of returns. For example, the take-up of the de-worming program studied by Miguel and Kremer (2004) was only 57%, despite the fact that the program was *free*, and that the only investment required was to sign an informed consent form (and some disutility for the child). Further, when a nominal fee was introduced in a randomly selected set of schools in the year after the initial experiment, the take-up fell by 80%, relative to free treatment (Kremer and Miguel (2003)). While this could be due to the fact that the private benefits are perceived to be low by the parents, it is worth noting that the hike in user fees happened after one year of free treatment, so that parents would have had time to observe the change in the child's health and attendance at school. Moreover, Kremer and Miguel (2003) also observe that, as long as the price was positive, there was no impact from the actual price on the take-up of the drug. This strong non-linearity between a price of zero and any positive price (which is also consistent with the evidence from school uniforms) appears to be inconsistent with a response to uniforms.

To sum up, the evidence suggests that, while investment seems to respond in part to the cost and the benefits of these investments, it appears to do so in ways that suggest that it does not only respond to returns as we are measuring them.

2.2.3 Taking Stock: Investment Rates

Investment rates, both in physical and human capital, are typically no higher in poorer countries than in rich countries. In part, at least, this probably reflects the fact that investment in poor countries does

not always respond to the availability of high returns.

3 Understanding Rates of Return and Investment Rates in Poor Countries: Aggregative Approaches

The aggregative approach does not aspire to explain why the marginal product varies so much within the same economy. From the point of view of this approach, the main puzzle, put forward for example in the Lucas calculation, is why the average rates of returns on investment in a poor country such as India are not higher compared to what they are in the U.S. Given that the differences in the rates of returns are relatively small, the lack of a bigger difference in investment rates is perhaps less of a mystery.

Though Lucas does not mention it, there is another, equally puzzling, observation lurking in the macro numbers. Given the existing capital stock, output-per-worker in India should be higher than what it is. To see this, recall from equation 4 that assuming that workers are only 30% as productive is equivalent to assuming that TFP in India should be approximately 50% of what it is in the U.S. This, combined with the fact that, according to the Penn World Tables, the U.S. has 18 times more capital-per-worker than India implies that the marginal product of capital ought to be $\frac{1}{2}(18)^{0.6} = 2.8$ times higher in India, which tells us that the marginal product in India ought to be about 25%, which is about what it is. However, using equation 3 in section 2.1.3, we calculated that the difference in output-per-worker between India and the U.S. implied that the rate of return to capital should be 45% in India.

The discrepancy between this number and the 25% we are now getting points to a second puzzle: It tells us that the differences in capital-per-worker that would be implied by the difference in output-per-worker is larger than the observed difference in capital-per-worker. Another way to see this is to substitute the numbers for capital-per-worker into the production function:

$$\frac{y_u}{y_i} = 2(\frac{k_u}{k_i})^{0.4} = 2 * (18)^{0.4} = 6.35. \tag{5}$$

This is obviously not nearly as large as the observed gap of 11:1. In other words, the second puzzle is to explain why output-per-worker in India is so low, given everything else, and in particular given the amount of capital in India.[11]

In the rest of this section we explore three different views of why one economy may lag behind another in terms of productivity. What they have in common is they operate entirely at the level of the economy, and do not directly suggest a theory of why there is so much inefficiency within a given economy.

[11]Note that this puzzle is *different* from the one Lucas started with, as we described earlier: The Lucas puzzle is simply the observation that, given the ratio in output-per-worker, the marginal product of capital is too low in India. Unlike the second puzzle, it makes no mention of the level of capital in India.

3.1 Access to Technology

One obvious answer to both puzzles is that TFP is not the same in India and the U.S. If TFP is lower, both productivity and the marginal product will be lower, for any given level of capital-per-worker.

One standard answer, within growth theory, of why TFP should be lower in poorer countries comes down to technology. There is a now a large literature—due to Aghion and Howitt (1992), Grossman and Helpman (1991) and others—that, emphasizes technological differences as the source of this TFP gap. It is easy to think of reasons why there may be a persistent technology gap between rich and poor countries. Essentially, it is too costly for the poor country to jump to the technological frontier because the frontier technologies belong to firms in the rich countries (who are the ones who have the biggest stake in keeping the technological frontier moving) and they charge monopolistic prices for access to these technologies.[12] Moreover, there is the issue of appropriate technology: The latest technology may not be suitable for use in a country with little human capital or poor infrastructure.

By itself, this explanation focuses on investment in technology and cannot directly account for the lack of investment in human capital in LDCs or why the returns there often seem so low. However, if there is strong complementarity between human capital investment and investments in new technology,[13] then the slow growth of TFP could explain the relative absence of investment in human capital in LDCs, assuming that we accept the rather mixed evidence, reviewed above, on the responsiveness of investment in human capital to the expected returns.

If the productivity gap between the U.S. and India has to be fully accounted for by technological differences in an aggregative model (i.e., if we rule out any differences in the interest rates), then TFP in the U.S. would have to be about twice that in India. How plausible is a TFP gap of 1:2 in a world of efficiently functioning markets? One way to look at this is to observe that U.S. TFP growth rates seem to be on the order of 1-1.5% a year. Even at 1.5%, TFP takes about 45 years to go up by 200%.[14] Therefore in 2000, Indians would have been using machines discarded by the U.S. in the 1950s.

This is also clearly very far from being true of the better Indian firms in most sectors. The McKinsey Global Institute's (McKinsey Global Institute (2001)) recent report on India, reports on a set of studies of the main sources of inefficiency in a range of industries in India in 1999, including apparel, dairy processing, automotive assembly, wheat milling, banking, steel, retail, etc. In a number of these cases

[12]The dominance of rich countries in the latest technologies is reinforced by the fact that the rich countries may have an actual advantage in R&D, because of their larger market size or their superior human capital endowments.

[13]As suggested, for example, in the work by Foster and Rosenzweig (1995) on the green revolution, which we discussed above.

[14]The effective rate of technological improvement will be larger, for example, if new technology needs to be embodied in machines and machines are more expensive (or savings rates are lower) in the poorer country (see Jovanovic and Rob (1997)).

(dairy processing, steel, software) they explicitly say that the better firms were using more or less the global best practice technologies wherever they were economically viable. The latest (or if not the latest, the relatively recent) technologies were thus both available in India and profitable (at least for some firms).

However, most firms do not make use of these technologies. And, according to the same McKinsey report, it is not because these technologies are not economically viable in this sector: The report on the apparel industry tells us that in the apparel industry:

> "Although machines such as the spreading machine provide major benefits to the production process and are viable even at current labor costs, they are extremely rare in domestic (i.e., non-exporting) factories." (McKinsey Global Institute (2001))

Despite this, technological backwardness is not one of the main sources of inefficiency that is highlighted in their report on the apparel industry. They focus, instead, on the fact that the scale of production is frequently too small, and in particular, on the fact that the median producer is a tailor who makes made-to-measure clothes at a very small scale, rather than a firm that mass produces clothes. TFP is low, not because the tailors are using the wrong technology, but because tailoring firms are too small to benefit from the best technologies and therefore should not exist.

Reports from a number of other industries show a similar pattern. Certain specific types of technological backwardness is mentioned as a source of inefficiency in both the dairy processing industry and the telecommunications industry, but in both cases it is argued that all firms should find it profitable to upgrade along these dimensions (McKinsey Global Institute (2001)).

In these two cases, however, there is also a reference to the gains (in terms of productive efficiency) from what the report calls "non-viable automation". This is automation that would raise labor productivity but lower profits. One reason why automation may be non-viable in this sense is that the technology may be under patent and therefore expensive, along the lines suggested by Aghion and Howitt (1992), Grossman and Helpman (1991) and others, or it may demand skills that the country does not have. Or, it could also be something entirely neo-classical: Labor-saving devices are less useful in labor-abundant countries. Since we have no way of determining why the technology is non-viable, we looked at the total labor productivity gain promised by this category of innovation. In both the dairy processing industry and the telecom case, this number is 15% or less, and in the automotive industry it is no larger (McKinsey Global Institute (2001)). This is clearly not large enough to explain the entire TFP gap.

On other hand, it is clearly true that there are many firms that, for some reason, have opted not to adopt the best practice despite the fact that others within the same economy find it profitable to do so and, at least according to McKinsey, they too would benefit from moving in this direction. If the

technological piece of the TFP gap turns out to be large, our presumption would be that it is driven by this second, more microeconomic, source of technological stickiness: Indeed, in their more recent work, Aghion and Mayer-Foulkes (2003) also emphasize the importance of the fact that firms may not have access to enough capital to implement the technologies that they would like to adopt.

3.2 Human Capital Externalities

Another source of difference between TFP that has been proposed (starting with Lucas), is the increasing returns stemming from human capital externalities: Human capital is valuable, not only in the firm that employs the worker, but for all firms.

This could explain a puzzle we did not discuss until now, pointed out by Acemoglu and Angrist (2001) and Bils and Klenow (2000): The high correlation between human capital and income that is observed in the cross-country data (e.g., Mankiw et al. (1992)) is hard to reconcile with the micro evidence we have reviewed earlier, which suggested relatively low returns to education. To see this, note that the difference in average schooling between the top and bottom deciles of the world education distribution in 1985 was less than 8 years. With a Mincerian returns to schooling of about 10%, the top decile countries should thus produce about twice as much per worker as the countries in the bottom decile. In fact, the output-per-worker gap is about 15. One possibility is that the Mincerian rate of return *understates* the true rate of returns to education, because it does not take into account positive externalities generated by educated workers. More specifically, the human capital externalities on the order of 20-25% (more than twice the private return) would be necessary to explain the cross-country relationship between education and income, which sounds implausible. Early evidence (e.g., Rauch (1993)) suggested that externalities were positive, but not of that order of magnitude. Using variation in education across U.S. cities, Rauch (1993) estimated that the human capital externalities may be on the order of 3% to 5%. Moreover, even this evidence is to be taken with caution, as cities where workers are more educated vary in many other respects. Using variation in average education generated by the passage of compulsory schooling laws, Acemoglu and Angrist (2001) find no evidence of average education on individual wages, after controlling for individual education.

In Indonesia, Duflo (2003) actually finds evidence that there may be *negative pecuniary* externalities across people who invest in education. She studies the impact of an education policy change that differentially affected different cohorts and different regions of Indonesia. Between 1973 and 1979, oil proceeds were used to construct over 61,000 primary schools throughout the country. Duflo (2001) shows that the program resulted in an increase of 0.3 years of education for the cohorts exposed to the program. Duflo (2003) takes advantage of the fact that individuals who were 12 or older when the program started did not benefit from the program, but worked in the same labor markets as those who did. As the

newly educated workers entered the labor force, starting in the 1980s, the fraction of educated workers in the labor force increased. Since migration flows in Indonesia remained relatively modest, the increase in the fraction of workers with primary education between 1986 and 1999 was faster in regions which received more INPRES schools. Using the interaction of year and region as instruments for the fraction of educated workers, she estimates that an increase of 10 percentage points in the fraction of educated workers in the labor force resulted in a *decrease* in the wages of the older workers (both educated and uneducated) by 4% to 10%. This suggests that, on balance, there are strongly diminishing aggregate returns at the local level: Any positive externality is more than compensated by these declining returns. The Mincerian returns could then actually *overestimate* the aggregate returns of increasing education, because by comparing individuals within a labor market, they do not take into account the diminishing returns that affect everybody in the labor market. In any case, at this point, there is no evidence that there are strongly increasing returns to human capital.

3.3 Coordination Failure

Another source of lower aggregate productivity is the possibility of coordination failures, which reduces aggregate productivity through a demand effect. There is a long line of work, starting with Rosenstein-Rodan (1943), that has emphasized the role of coordination failure in explaining why certain countries successfully industrialize, while others remain poor and non-industrialized. Murphy and Vishny (1989) explore models where industrialization in a sector creates demand for the products of another sector (through higher wages for the workers), and which leads to multiple equilibria. A coordinated "big push", where all industries start together, can place the country on a permanently higher level of investment and income. Developing countries may have low investments and low returns to capital because such a "big push" has not happened. A large literature explores different forms of strategic complementarities. Since the argument involves an entire economy's coordination, it is difficult to use micro-evidence to provide much direct evidence about these aggregate externalities.[15] However, while these theories certainly have some relevance, it is not entirely clear whether aggregate demand effects can be so powerful as to generate the necessary gap in TFP between, say, India and the U.S., given the existence of international trade. Further, this aggregate approach cannot explain the fact, reviewed earlier, of why some firms seem to adopt the latest technologies, while others do not, and why the marginal product of capital varies so much.

[15] Below, we will review the evidence on more local externalities.

3.4 Taking Stock

While the evidence is somewhat impressionistic, it seems unlikely that the aggregative theories discussed above can explain the entire TFP gap. Of course, if we were prepared to give up the idea that the entire problem comes from a lower aggregate productivity, for example by accepting that the marginal product is higher in India, the problem of fitting the data would be easier. For example, if the TFP gap were 1.5 times higher in the U.S. (on top of what is predicted by the difference in the productivity of labor), the fact that the U.S. has 18 times more capital-per-worker would imply that output-per-worker would be $(1.5)(2)(18)^{0.4} = 9.5$ times higher in the U.S., and the marginal product of capital would be $\frac{(18)^{0.6}}{2(1.5)} = 1.9$ times higher in India. These are both clearly in the ballpark, although the output gap between the U.S. and India predicted by this model is still too low (the output gap is about 11:1 in the data) and the ratio of the marginal product of capital between India and the U.S., which was too high in a model with identical TFP, is now too low (the ratio in the data is about 2.5).

It is worth noting that in order to get closer to the 11:1 ratio in the data, the TFP ratio would need to be higher than 1.5, which is perhaps already too big. Moreover, this would further reduce the predicted ratio between the marginal product of capital in India and in the U.S., which was already too low when the TFP gap was 1.5. In other words, we are facing a new problem: Given the existing capital stock, if a difference in TFP were the reason why the output-per-worker is so low in India, the marginal product of capital should be even lower in India than what it is. Indeed, there is no way to adjust the TFP ratio to improve the fit along both dimensions—we can increase the gap in output-per-worker by raising the TFP ratio, but only at the cost of making the ratio of marginal product even smaller. The problem is quite basic: With a Cobb-Douglas production function, the average product of capital is proportional to its marginal product. But then output-per-worker must be proportional to the product of the marginal product of capital and capital-per-worker. If the marginal product in India is 2.5 times that of the U.S., but capital-per-worker is 18 times greater in the U.S., output-per-worker has to be $\frac{18}{2.5} = 7.2$ times larger in the U.S. and not 11 times larger, *irrespective of what we assume about the ratio of TFP in the two countries.* In other words, the only way we can hope to really fit what we see in the data is by abandoning the standard Cobb-Douglas formulation. This is useful to keep in mind when, in later sections, we discuss ways to improve the fit between the theory and the data.

To sum up, Lucas' question about why capital does not flow from the U.S. to India was, in some sense, where it all started, but from the vantage point of what we know today, this is in some ways the lesser problem. We know now that there are differences in the marginal product of capital within the same economy that dwarf the gap that Lucas calculated from comparison of India and the U.S., and found so implausibly large that he set out to rewrite all of growth theory. The harder question is why capital flows do not eliminate these differences.

Lucas' resolution of the puzzle was to give up the key neo-classical postulate of equal TFP across countries. Based on the McKinsey report, this seems to be the obvious step, but the problem is less that people in developing countries do not find it profitable to adopt the latest (and best) technologies and more that many firms do not adopt technologies that are available and would be profitable if adopted. The key question, once again, is why the market allows this to be the case.

The premise of the aggregative approach to growth was that markets function well enough within countries that we can largely ignore the fact that there is inefficiency and unequal access to resources within an economy when we are interested in dynamics at the country level. The evidence suggests that this is not true: The cross-country differences in marginal products or technology that we want to explain are of the same order of magnitude as the differences we observe within each economy. Development economists are therefore more interested in theories of cross-country differences that also help us understand why rates of returns vary so much within each country. This is what we turn to next: In section 4, we first review the various reasons that have been proposed. In section 5, we will then calibrate their impact to evaluate whether they can form the basis of an explanation for the puzzles we observed.

4 Understanding Rates of Return and Investment Rates in Poor Countries: Non- Aggregative Approaches

In this section, we review various possible reasons why individuals do not always make the best possible use of resources available to them.

4.1 Government Failure

One reason why firms may not choose the latest technologies or make the right investments is because they do not have the proper incentives to do so. A line of work has developed the hypothesis that governments are largely responsible for this situation, either by not protecting investors well enough or by protecting some of them excessively. The firms that are ill-protected underinvest and have high marginal returns, while the over-protected firms overinvest and show low marginal returns. The net effect on investment may be negative, because even those who are currently favored may fear a future falling out and a corresponding loss of protection. Overall productivity may also go down, since the right people may not always end up in the right business, since connections rather than skills will dominate the choice of professions.[16]

One approach to investigating this hypothesis has been to try to document variations in the quality

[16] See Murphy and Vishny (1995).

of institutions, and to try to evaluate their impact. La Porta and Vishny (1998) document important variations in the degree to which the law protects investors (creditors and shareholders) across countries, part of which seem to be explained by the origin of these countries' legal codes (the French civil law has much less legal protection for investors than Anglo-Saxon common law). Djankov and Shleifer (2002) document wide variation in the ability of someone to start a new firm in 85 countries. They argue that the costs of entry are high in most countries (on average, they sum up to 47% of a country's GDP per capita), and can be very high indeed: While it take 3 procedures and 3 days to obtain the permit to start company in New Zealand, it takes 19 procedures, 149 business days and 111.5 percent of GDP per capita in Mozambique. The procedure is shorter, and generally less expensive in terms of GDP per capita, in rich countries than in poor or middle-income countries. Djankov and Shleifer (2003) document the time it takes in court to evict a tenant or collect a bounced check, as well as the degree of formalism of the legal procedures. They find, once again, wide variation: In particular, these procedures take a much shorter time in countries with common law legal origins. Similarly, many studies argue that, in cross-country regressions, there is a strong association between aggregate investment and measures of bad institutions or corruption (e.g., Knack and Keefer (1995), Mauro (1995), Svensson (1998)).

These papers also argue that low investor protections, legal barriers to entry, and long legal procedures have implications for welfare and efficiency. There are indeed suggestive associations in the data (for example, ownership is more concentrated when investor protection is worst), but there is always the possibility that the correlation between the quality of the institutions and the real outcomes they consider is due to a third factor. Acemoglu and Robinson (2001) try to address this issue by finding exogenous variations in the quality of institutions. They argue that there is a persistence of institutions, so that countries which accessed independence with extractive institutions (e.g., Congo) have tended to keep these bad institutions. They then argue that colonial powers were more likely to set up extractive institutions, with an unrestrained executive power, in places where they did not intend to settle. Finally, they were less likely to settle in places where the environment was hostile: In particular, the mortality of early settlers predicted the number of people of European descent who settled in these countries, the quality of institutions at the turn of the 20th century, and the quality of institutions in 1995 (measured as the risk of expropriation perceived by investors). In turn, it also is associated with lower GDP in 1995. The authors then use early settler mortality as an instrument for institutions in a regression of the impact of institutions on inequality, and find a strong positive coefficient.

This evidence suggests that government matters, and that bad government will lower returns and discourage new investments. There is a literature that tries to investigate the exact mechanisms through which the government affects the allocation of resources. One version of the story blames excessive intervention, while another talks about the lack of appropriate regulations. We now discuss these two

explanations in turn, and try to assess how far they can help us fit the evidence.

4.1.1 Excessive Intervention

There is a line of work, following Parente and Prescott (1994, 2000), which argues that the productivity gap results from the way the heavy hand of the government operates. The government makes rules that discourage innovation and protects the inept, and thereby slows the economy's progress towards the ideal state where only the most productive firms survive.

There is clearly something to this vision. Gelos and Werner (2002) show that financial de-regulation in Mexico (which started in 1988 and eliminated the interest rate ceiling, high reserve requirements which channeled 72% of commercial bank lending to the government, and priority lending) increased the ability of small firms to access the credit market, and reduced the excess cash flow sensitivity of investment *for small firms only*. Until recently in India, a large number of sectors were reserved for firms below a certain size (the small-scale sector) and/or firms in the cooperative sector. Small firms also benefited from tax exemptions and priority sector credits. This clearly limited the ability to take advantage of economies of scale and restricted the market share of the most efficient players.

Nonetheless, this is probably only a part of the story. As we noted in the context of the discussion of Banerjee and Duflo (2003a), even medium-sized firms that were well above the cut-off for being included in the small-scale sector seem to be operating well below their optimal scale. In other words, notwithstanding the politically protected presence of the small-scale firms that is presumably driving down profits in the sector, these medium-sized firms were clearly still at the point where further investment would be extremely profitable. There has to be something other than a policy-induced lack of profitability that was holding them back.

The same point is made in a different style in the paper by Banerjee and Munshi (2004), mentioned above. This paper studies investment and productivity differences among firms in the knitted-garment industry in Tirupur, India. The firms owned by the Gounders, tend to be much larger than the firms owned by all other participants in the industry: The gap among firms that had just started is on the order of three to one. Yet these Gounder firms produce much less per unit of capital, and Gounder firms that have been in business for more than five years actually produce less *in absolute terms* than the smaller firms of the same vintage owned by non-Gounders. In other words, it is the bigger firms that are less productive, in an environment where the government discriminates, if at all, in favor of the smaller firms.

To sum up, while there are certainly instances of excessive intervention, it seems that there are many inefficiencies that cannot be blamed on the government.

4.1.2 Lack of Appropriate Regulations: Property Rights and Legal Enforcement

Effective rates of return and investment rates can be low because the responsibilities and/or the benefits of the investments are shared, or the investors are worried about being expropriated: The investor is therefore not capturing the full marginal returns of its investment. Imperfect property rights will thus lead to low investments. Poorly enforced property rights also make it difficult to provide collateral, which exacerbates the problems of the credit market. For example, the study of the Mexican financial deregulation discussed above (Gelos and Werner (2002)) showed that after the deregulation, small firms' access to credit became more linked to the value of the real estate assets they could use as collateral: The role of the government does not end with not interfering, it may also be to provide secure property rights.

In addition to the macro-economic evidence mentioned above, there is some micro-economic evidence that property rights matter for investment, although the findings are more mixed. Goldstein and Udry (2002) show that, in Ghana, individuals are less likely to leave their land fallow (which is an investment in long run land productivity) if they do not hold a position of power within the family of the village hierarchy which ensures that their land is not taken away from them when it is fallow. However, Besley (1995) finds that, also in Ghana, investment (tree planting) is not significantly larger when individuals have more secure rights to their land. Johnson and Woodruff (2002) find that, in five post-Soviet countries, firms that are run by entrepreneurs who perceive that their property rights are more secure invest more than those who do not. The effect is as strong for firms who rely mostly on internal finances as for those who have access to external finance. Entrepreneurs who believe that they have strong property rights invest 56% of their profits in their firms (against 32% for those who do not). Do and Iyer (2003) find that a land reform which gave farmers the right to sell, transfer or inherit their land usage rights also increased agricultural investment, in particular the planting of multi-year crops (such as coffee).

Even when property rights themselves are legally well defined and protected, there are institutions which reduce the private incentives to invest. Sharecropping is one environment where both the landlord and the tenants have low incentive to invest in the inputs that they are responsible for providing (Eswaran and Kotwal (1985)). Binswanger and Rosenzweig (1986) and Shaban (1987) both show that, controlling for farmer's fixed effect (that is, comparing the productivity of owner-cultivated and farmed land for farmers who cultivate both their own land and that of others) and for land characteristics, productivity is 30% lower in sharecropped plots. Shaban (1987) shows that all the inputs are lower on sharecropped land, including short-term investments (fertilizer and seeds). He also finds systematic differences in land quality (owner-cultivated has a higher price per hectare), which could in part reflect long-term investment. Banerjee and Ghatak (2002) study a tenancy reform which increased the tenants' bargaining power and security of tenure. They found that the land reform resulted in a substantial increase in the productivity

of the land (62%). Since the reform took place at the same time as the green revolution, this increase in productivity is probably in part due to an increased willingness to switch to the new seeds after the registration program.[17]

The example of sharecropping suggests that bad governments are not the only cause for the emergence of bad institutions. If sharecropping is inefficient, why does it arise? In particular, why do the landlord and the tenant not agree on a fixed rent, which will ensure that the tenant is the full beneficiary of his effort at the margin? Explanations of the persistence of sharecropping involve risk aversion (Stiglitz (1974)) or limited liability (Banerjee and Ghatak (2002)). This suggests that while the proximate explanation for inefficient investment may well be based in a specific institution, the more basic cause may be lying elsewhere, in the way various asset markets function. This is what we turn to next.

4.2 Credit Constraints

• Why would credit markets function poorly in poor countries?

The fact that the capital market does not function well in poor countries is a result of a number of factors. First, information systems, including property records, are often underdeveloped, making it hard to enforce contracts. This, in turn, partly reflects the fact that people may not know how to read or write and partly the fact that there has not been enough institutional investment.[18] Second, the fact that potential borrowers are poor and under extreme economic pressure, might make them all too willing to try to cheat the lender. Third, there are political pressures to protect borrowers from lenders in most LDCs.

• Consequence of poorly functioning credit market.

Given the problems in enforcing the credit contract, what a lender will be prepared to offer a particular borrower will depend on the quality of the borrower's collateral, his reputation in the market, the ease of keeping an eye on him and a host of other characteristics of the borrower. This has the obvious implication that two firms facing the exact same technological options may end up choosing very different methods of production. In particular, one person may start a large or more technologically advanced firm because he has money and another may start a small and backward one because he does not. As a result, neither interest rates, nor TFP, nor the marginal product need be equalized across borrowers.

This would also explain why investment responds so unpredictably to returns: Sometimes the oppor-

[17]This interpretation is reinforced by the fact that their estimates are higher than Shaban's and that of a study by Laffont and Matoussi (1995) who use data from Tunisia to show that a shift from sharecropping to fixed-rent tenancy or owner cultivation raised output by 33 percent, and moving from a short-term tenancy contract to a longer-term contract increased output by 27.5 percent.

[18]For example, Djankov and Shleifer (2002) document the time it takes the recover a bounced check across countries. It takes longer in poorer countries.

tunities become available when there is large group of people who are looking to invest and have the wherewithal to do it. At other times, the returns may be there but most of those who have money may be heavily involved in promoting something else.

A second set of implications of imperfect contracting in the credit market is that the supply curve of capital to the individual borrower slopes up—a borrower who is more leveraged will need more monitoring and the lender will charge him more to do the extra monitoring. And eventually, the extra monitoring may be too costly to be worth it, and the borrower will face an absolute limit on how much he can borrow.

An immediate consequence of an upward-sloping supply curve is that the marginal product of capital will be higher than what the borrower pays the lender. Indeed, the gap between the two may quite substantial, since the fact that borrowers are constrained in borrowing also implies that the lenders are constrained in how much they can lend at rewarding rates. This drives the interests rates down, as lenders compete for the best borrowers. Moreover, since the rates the lenders charge include the cost of the monitoring that they have to do, the rates the lenders charge could be much higher than the opportunity cost of capital. In the case of a financial intermediary, such as a bank, this implies that the rates they charge their borrowers may be much higher than the rates they pay their depositors.

This implies, for example, that the American investor who gets 9% on his stock market investment could not just put the money in a bank in India and earn the 22.5% average marginal product. Indeed, he may not earn much more than 9% if he were to put it in an Indian bank. However, he could set up a business in India and earn those returns, and presumably if enough people did that, the returns would be equalized; below we will try to say something about why this does not happen.

It also implies that the incentive to save may be low in countries where the marginal product is high, except for those who are planning to invest directly. This might help to explain the low equilibrium investment rate, though it is theoretically possible that the negative effect on the savers would be swamped by the positive effect on investors if the fraction of investors is large enough.

• Evidence

We have already mentioned some evidence from South Asia showing that the interest rate varies enormously across borrowers within the same local capital market and that the extent of variation is too large to be explained by the observed differences in default rates. Banerjee (2003a) lists a number of studies that make it clear that this is also true in developing countries outside South Asia. This is suggestive, albeit indirect, evidence of credit constraints.

If the marginal product of capital in the firm is greater than the market interest rate, credit constraints naturally mean that a firm would want to borrow more than what is available. It is, however, not clear how one should go about estimating the marginal product of capital. The most obvious approach, which relies on using shocks to the market supply curve of capital to estimate the demand curve, is only valid under

the assumption that the supply is always equal to demand, i.e., if the firm is never credit constrained.

The literature has therefore taken a less direct route: The idea is to study the effects of access to what are taken to be close substitutes for credit—current cash flow, parental wealth, community wealth—on investment. If there are no credit constraints, greater access to a substitute for credit would be irrelevant for the investment decision. While this literature has typically found that these credit substitutes do affect investment,[19] suggesting that firms are indeed credit constrained, the interpretation of this evidence is not uncontroversial. The problem is that access to these other resources is unlikely to be entirely uncorrelated with other characteristics of the firm (such as productivity) that may influence how much it wants to invest. To take an obvious example, a shock to cash-flow potentially contains information about the firm's future performance.

The estimation of the effects of credit constraints on farmers is significantly more straightforward since variation in the weather provides a powerful source of exogenous short-term variation in cash flow. Rosenzweig and Wolpin (1993) use this strategy to study the effect of credit constraints on investment in bullocks in rural India.

The paper by Banerjee and Duflo (2003a) that we discussed above makes use of an exogenous policy change that affected the flow of directed credit to an identifiable subset of firms in India. Since the credit was subsidized, an increase in sales and investment as a response to the increase in funds available needs to mean that firms are credit constrained, since it may have decreased the marginal cost of capital faced by the firm. However, they argue that if a firm is not credit constrained then an increase in the supply of subsidized directed credit to the firm must lead it to substitute directed credit for credit from the market. Second, while investment, and therefore total production, may go up even if the firm is not credit constrained, it will only go up if the firm has already fully substituted market credit with directed credit. They showed that bank lending and firm revenues went up for the newly targeted firms in the year of the reform. They find no evidence that this was accompanied by substitution of bank credit for borrowing from the market and no evidence that revenue growth was confined to firms that had fully substituted bank credit for market borrowing. As already argued, the last two observations are inconsistent with the firms being unconstrained in their market borrowing.

The logic of credit constraints applies as much or more to human capital investments. Hart and Moore (1994), among others, have used human capital as the archetype of investment that cannot be collateralized, and therefore is hard to borrow against. This is made even more difficult by the fact that children would need to borrow for their education, or parents would need to borrow on their behalf. We

[19]The literature on the effects of cash-flow on investment is enormous. Fazzari and Petersen (1988) provide a useful introduction to this literature. The effects of family wealth on investment have also been extensively studied (see Blanchflower and Oswald (1998) for an interesting example). There is also a growing literature on the effects of community ties on investment (see, for example, Banerjee and Munshi (2004)).

return to this evidence below. The high responsiveness to user fees that we reviewed in section 2, and the evidence that investment in education are sensitive to parental income,[20] are both consistent with credit constraints. However, because human capital investments may involve direct utility or disutility (for example, a parent may like to see his child being educated), it is more difficult to come up with evidence that systematically nails the role of credit constraints for human capital investment. Edmonds (2004) is an interesting attempt to try to isolate the effect of credit constraints using household's response to an *anticipated* income shock. He studies the effect on child labor and education of a large old age pension program, introduced in South Africa at the end of the Apartheid. Many children live with older family members (often their grandparents). Women become eligible at age 60 and men become eligible at age 65. Since at the time he studies the program, the program was well in place and therefore fully anticipated, he argues that if more children start attending school as soon as their grandfather or grandmother crosses the age threshold and becomes eligible (rather than continuously, as they come closer to eligibility), this must be an indication of credit constraint. Indeed, he finds that child labor declines, and school enrollment increases, discretely when a household member becomes eligible.

• Summary

Credit constraints seem to be pervasive in developing countries. Of course, we are interested in whether the fact that access to capital varies across people helps us understand the productivity gap. If people invest different amounts because of differential access to capital, our intuitive presumption would be that capital is being misallocated, because there is no reason why richer people are always better at making use of the capital. This misallocation could be a source of difference in productivity. We will return to this question in section 5.

4.3 Insurance Market Failures

Even if credit markets function well, and there is no limited liability, individuals may be reluctant to invest in any risky activity, for fear of losing their investment, if they are not properly insured against fluctuations in their incomes. Risk aversion leads to inefficient investment, and efficiency would improve with insurance (this idea is explored theoretically in Stiglitz (1969), Kanbur (1979), Kihlstrom and Laffont (1979), Banerjee and Newman (1991), Newman (1995) and Banerjee (2001)).

• Insurance in developing countries.

A considerable literature has investigated the extent of insurance in rural areas in developing countries (see Bardhan and Udry (1999) for a survey). Townsend (1994) used the ICRISAT data, a very detailed panel data set covering agricultural households in four villages in rural India to test for perfect insurance. The main idea behind this test is that with perfect insurance at the village level only aggregate

[20] See Strauss and Thomas (1995) for several studies along these lines.

(village-level) income fluctuation, and not idiosyncratic income fluctuations, should translate into fluctuation in individual consumption. He was unable to reject the hypothesis that the villagers insure each other to a considerable extent: Individual consumption seems to appear to be much less volatile than individual income, and to be uncorrelated with variations in income. This exercise had limits, however (see Ravallion and Chaudhuri (1997) for a comment on the original paper), and subsequent analyses, notably by Townsend himself, have shown the picture to be considerably more nuanced. Deaton (1997) shows that there is no evidence of insurance in Cote d'Ivoire. Townsend (1995) finds the same results across different areas in Thailand. Fafchamps and Lund (2003) find that, in the Philippines, households are much better insured against some shocks than against others. In particular, they seem to be poorly insured against health risk, a finding corroborated by Gertler and Gruber (2002) in Indonesia. Most interestingly, Townsend (1995) describes in detail how insurance arrangements differ across villages. While in one village there is a web of well-functioning risk-sharing institutions, the situations in other villages are different: In one village, the institutions exist but are dysfunctional; in another village, they are non-existent; finally, in a third village, close to the roads, there seems to be no risk-sharing whatsoever, even within family.

This last fact is attributed to the proximity to the city, which makes the village a less close-knit community, where enforcement of informal insurance contracts is more difficult. Coate and Ravallion (1993) was the first paper to build a theoretical model of insurance with limited commitment, and to show that, when the only incentive to contribute to the insurance scheme in good times is the fear of being cut away from the insurance in future periods, insurance will be limited. It will also be optimal to make payment contingent on past history, which will lead to a blur between credit and insurance (Ray (1998)). Udry (1990) presents evidence from Nigeria that is consistent with this model. The villages he studies are characterized by a dense network of loan exchange: Over the course of one year, 75% of the households had made loans, 65% had borrowed money, and 50% had been both borrowers and lenders. Ninety-seven percent of these loans took place between neighbors and relatives. Most importantly, the loans are "state-contingent": Both the repayment schedule and the amount repaid are affected by the lender's state *and* the borrower's state. This is evidence that credit is to some extent used as an insurance device. The resulting system is a mix of credit and insurance close to what the model of limited commitment would predict. However, and still consistent with this prediction, there is not enough of this "security" to fully insure households against income fluctuations: A shock to a particular borrower has a negative impact on the sum of the transfers received by his lender, which means that the lender did not fully diversify risk.

Despite this evidence, we do not fully understand the reasons for the lack of insurance among households. It is unlikely that either limited commitment or the more traditional explanations in terms of

moral hazard or adverse selection can explain why the level of insurance seems to vary from one village to the next, or why there is no more insurance against rainfall, for example.

• Consequences for investment.

Irrespective of the ultimate reason for the lack of insurance, it may lead households to use productive assets as buffer stocks and consumption smoothing devices, which would be a cause for inefficient investment. Rosenzweig and Wolpin (1993) argue that bullocks (which are an essential productive asset in agriculture) serve this purpose in rural India. Using the ICRISAT data, covering three villages in semi-arid areas in India, they show that bullocks, which constitute a large part of the households' liquid wealth (50% for the poorest farmers), are bought and sold quite frequently (86% of households had either bought or sold a bullock in the previous year, and a third of the household-year observations are characterized by a purchase or sale), and that sales tend to take place when profit realizations are high, while purchases take place when profit realizations are low. Since there is very little transaction in land, this suggests that bullocks are used for consumption smoothing. Because everybody needs bullocks around the same time, and bullocks are hard to rent out, Rosenzweig and Wolpin estimate that, in order to maximize production efficiency, each household should own exactly two bullocks at any given point in time. The data suggest that, for poor or mid-size farmers there is considerable underinvestment in bullocks, presumably because of the borrowing constraints and the inability to borrow and accumulate financial assets to smooth consumption: Almost half the households in any given year hold no bullock (most of the others own exactly two).[21] Using the estimates derived from a structural model where household use bullocks as a consumption smoothing device in an environment where bullocks cannot be rented and there is no financial asset available to smooth consumption, they simulate a policy in which the farmers are given a certain non-farm income of 500 rupees (which represents 20% of the mean household food consumption) every period. This policy would raise the average bullock holding to 1.56, and considerably reduce its variability, due to two effects: The income is less variable, and by increasing the income, it makes "prudent" farmers (farmers with declining absolute risk aversion) more willing to bear the agricultural risk.

Moreover, we observe only insurance against the risks that people have chosen to bear; the inability to smooth consumption against variation in income may lead households to choose technologies that are less efficient, but also less risky. Banerjee and Newman (1991) argue, for example, that the availability of insurance in one location (the village), while its unavailability in another (the city), may lead to inefficient migration decisions, since some individuals with high potential in the city may prefer to stay in the village

[21] The fact that there is under-investment on average, and not only a set of people with too many bullocks and a set of people with too few, is probably due to the fact that bullocks are a lumpy investment, and owning more than two is very inefficient for production–there is no small adjustment possible at the margin.

to remain insured.

There is empirical evidence that households' investment is affected by the lack of *ex post* insurance. Rosenzweig and Binswanger (1993) estimate profit functions for the ICRISAT villages, and look at how input choices are affected by variability in rainfall. They show that more variable rainfall affects input choices, and in particular, poor farmers make less efficient input choices in a risky environment. Specifically, a one standard deviation increase in the coefficient of variation of rainfall leads to a 35% reduction in the profit of poor farmers, 15% reduction in the profit of median farmers, and no reduction in the profit of rich farmers. Morduch (1993) specifically investigates how the anticipation of credit constraint affects the decision to invest in HYV seeds. Using a methodology inspired by Zeldes (1989), he splits the sample into two groups, one group of landholders who are expected to have the ability to smooth their consumption, and one group that owns little land, whom we expect *a priori* to be constrained. He finds that the more constrained group uses significantly less HYV seeds.

It is worth noting that the estimated impact of lack of insurance on investment is likely to be a serious underestimate. It is not clear how one could evaluate how much the lack of insurance affects investment. While we might observe certain options considered by the investor, there is no obvious way for knowing what other, even more lucrative choices, he chose not to even think about.

4.4 Local externalities

As we discussed in section 4, there is a line of work that focuses on coordination failures at the level of the economy: However, Durlauf (1993) shows that externalities do not have to be aggregated for the economy to exhibit multiple equilibria: Local complementarities (where adoption of a particular technology lowers production costs in a few "neighboring" sectors) can build up over time to affect aggregate behavior and generate lower aggregate growth.

An example of strategic complementarity of this kind arises when agents are learning from each other. Banerjee (1992) shows how, when people try to infer the truth from other people's actions, this leads them to under-utilize their own information, and leads to "herd behavior". While this behavior is rational from the point of view of the individual, the resulting equilibrium is inefficient, and can lead to underinvestment, overinvestment, or investment in the wrong technology whatsoever.[22]

The impact of learning on technology adoption in agriculture has been studied particularly extensively. Besley and Case (1994) show that in India, adoption of HYV seeds by an individual is correlated with adoption among their neighbors. While this could be due to social learning, it could also be the case that common unobservable variables affect adoption of both the neighbors.[23] To partially address this problem,

[22]For a related model, see Bikhchandani, Hirshleifer and Welch (1992).

[23]See Manski (1993) for a discussion of the identification problem in social learning problems.

Foster and Rosenzweig (1995) focus on profitability. As we mentioned previously, during the early years of the green revolution, returns to HYV were uncertain and dependent on adequate use of fertilizer. In this context, the paper shows that profitability of HYV seeds increased with past experimentation, by either the farmers or others in the village. Farmers do not fully take this externality into account, and there is therefore underinvestment. In this environment, the diffusion of a new technology will be slow if one neighbors' outcomes are not informative about an individual's own conditions.[24] Indeed, Munshi (2003) shows that in India, HYV rice, which is characterized by much more varied conditions, displayed much less social learning than HYV wheat.

All of these results could still be biased in the presence of spatially correlated profitability shocks. Using detailed information about social interactions Conley and Udry (2003) distinguish geographical neighbors from "information neighbors", the set of individuals from whom an individual neighbor may learn about agriculture. They show that pineapple farmers in Ghana imitate the choices (of fertilizer quantity) of their information neighbors when these neighbors have a good shock, and move further away from these decisions when they have a bad shock. Conley and Udry try to rule out that this pattern is due to correlated shocks by observing that the choices made on an established crop (maize-cassava intercropping), for which there should be no learning, do not exhibit the same pattern.

The ideal experiment to identify social learning is to exogenously affect the choice of technology of a group of farmers and to follow subsequent adoption by themselves and their neighbors, or agricultural contacts. Duflo and Robinson (2003) performed such an experiment in Western Kenya, where less than 15% of the farmers use fertilizer on their maize crop (the main staple) in any given year despite the official recommendation (based on results from trials in experimental farms), as well as the high returns (in excess of 100%) that they estimated. They randomly selected a group of farmers and provided fertilizer and hybrid seeds sufficient for small demonstration plots in these farmers' fields. Field officers from an NGO working in the area guided the farmers throughout the trial, which was concluded by a debriefing session. In the next season, the adoption of fertilizer by these farmers increased by 17%, compared to the adoption of the comparison group. However, there is no evidence of any diffusion: People named by the treatment farmers as people they talk to about agriculture did not adopt fertilizer any more than the contacts of the comparison group. The neighbors of the treatment group actually tended to adopt fertilizer *less* often, relative to the neighbors of the comparison group. This is not because only experimentation in one's own field changes someone's priors: When randomly selected friends were invited to attend the harvest, the debriefing session, and other key periods of the trials, they were as likely to adopt fertilizer as the farmers who participated in the experiment. Rather, it suggests that, spontaneously, information

[24] Ellison and Fudenberg (1993) describe "rule of thumb" learning rules where individuals learn from others only if they are similar.

about agriculture is not shared. This points towards another type of externality and source of multiple equilibria: When there is very little innovation in a sector, there is no news to exchange, and people do not discuss agriculture. As a result, innovation dies out before spreading, and no innovation survives.

Depending on the priors of the individuals, social learning can either decrease or increase investment. In Kenya, Miguel and Kremer show that random variation in the number of friends of a child who was given the deworming medicine had a *negative* impact of the propensity of a child to take the medicine. They attribute this to the fact that parents may have initially over-estimated the benefits of the deworming drug.

In addition to social learning, there are many other sources of local interactions. First, people imitate each other even when they are not trying to learn, because of fashion or social pressure. Social norms may prevent the adoption of new technologies, because coordinating on a new equilibrium may require many people to change their practices at the same time.[25] Second, there are several sources of positive spillovers between industries located close to each other. Silicon Valley-style geographic agglomerations occur in the developing world as well, such as the software industry in Bangalore. Ellison and Glaeser (1997) show that, in the U.S., most industries are indeed more concentrated than they would be if firms decided to place their plants randomly. Only about half of this concentration is explained by the fact that some locations have natural advantages for (Ellison and Glaeser (1999)) specific industries.

In addition to the traditional arguments for positive spillovers, such as transport costs (fast telecommunication lines that were installed for the software industry in Bangalore greatly reduced the cost of setting up call centers, for example), intellectual spillovers or labor market pooling, a powerful reason for geographical agglomeration in developing countries is the role of a town's reputation in the world market. For example, outsiders who want to start working in garment manufacturing come to Tirupur, the small town studied in Banerjee and Munshi (2004), despite their difficulty of finding credit there, because this is the place where large American stores come to place orders. There is a sense in which the town has a good reputation, for quality and timeliness of the delivery, and everybody who works there benefits from it. Tirole (1996) models "collective reputation": If many people in a group are known to deliver good quality products, buyers will have high expectations and be willing to trust the sellers to produce more elaborate products, where quality matters. In turn, this will encourage sellers to produce high quality products to avoid being outcast from the group, which will sustain a "high quality-high trust equilibrium". But if buyers are expected to only ask for basic products in the future, building a reputation for high quality is not useful, and opportunistic sellers will produce low quality in the first period. Knowing this, sellers indeed have the incentive to ask for simple products, and the bad equilibrium persists. In this world, history matters. A collective reputation for low quality is very difficult to

[25] See Munshi and Myaux (2002) for an example on the spread of family planning in Bangladesh.

reverse, and a collective reputation for high quality is valuable. We should therefore expect groups to try to set up institutions to develop a good collective reputation. There is certainly some indication that this is happening. For example, the association of Indian software firms (NASSCOM) tries to help the firms access quality certifications such as ISO 9001, SEI, or others. Much more work on whether collective reputation matters in practice is, however, clearly needed before we can assess the empirical relevance of these sources of externality.

To summarize, externalities can explain very large variations in productivity and investment rates across otherwise similar environments.

4.5 The Family: Incomplete Contracts Within and Across Generations

Investment in human capital often pays in the long term, and in many crucial instances must be done by parents on behalf of the child. In this context, the way the decisions are made in the family has a direct impact on investment decisions. In the benchmark neo-classical model (Barro (1974); Becker (1981)), parents value the utility of their children, perhaps at some discounted rate. This world tends to be observationally equivalent to one where an individual maximizes his long run income, and has the same strong convergence properties. However, if parents are not perfectly altruistic, the ability to constrain the repayment of future generations influences investment decisions. Banerjee (2003b) studies the short and long run implications of different ways to model the family decision-making process. He shows that incomplete contracting between generations generates potentially large deviations from the very strong convergence property of the Barro-Becker model. Deviations also occur if parents value human capital investment for its own sake (for example, because people like to see their children happy).[26]

In particular, even with perfect credit markets, parental wealth will determine how much is invested in human capital. There can be more than one steady state, and there can be inequality in equilibrium. In this world, increases in returns to human capital may not lead to an increase in human capital, if the production of human capital is skill-intensive (the increase in the price of teachers may dominate the added incentives to invest in education).

Many studies have shown that human capital investment is correlated with family income (see Strauss and Thomas (1995) for references for developing countries). In general, however, it is difficult to separate out the pure income effect from the effect of an increase in the returns to investing in human capital,

[26]Part of the reason why investment in human capital may appear like a preference factor is that individuals want their offspring to thrive and survive. In the U.S., Case and Paxson (2001) and Case and McLanahan (2000) find that investment in children is lower when they do not live with their birth mother. Using data from several African countries, Case and Ableidinger (2002) find that the gap between the probably of being enrolled in school for orphans and non-orphans can be in part accounted for by the fact that they are less likely to live with at least one parent, and more likely to live with non-relatives.

differences in the opportunity cost or the direct cost of schooling, and different discount rates. For example, in the Barro-Becker model, families with a lower discount rate will tend to be richer and more likely to invest in education. To avoid this problem, a few studies have focused on exogenous changes in government transfers. For example, Carvalho (2000) shows that an increase in pension income in Brazil led to a decrease in child labor and an increase in school enrollment. Duflo (2003) shows that, in South Africa, girls (though not boys) have better nutritional status (they are taller and heavier) in households where a grandmother is the recipient of a generous old age pension program.

This paper also touches on another set of issues. Different members of the family may have different preferences. If education and health were pure investment, and if the members of the household bargained efficiently (as in Lundberg and Pollack (1994, 1996) or the papers reviewed in Bourguignon and Chiappori (1992)), this would not have any impact on education or health decisions. However, if either assumption is violated, it means that not only the size of the income effects, but who gets the income, will affect investment decisions. In the case of the South African pensions, this was clearly the case: Pensions received by men had no impact on the nutritional status of children of either gender. This may come from the fact that women and men value child health differently, or from the fact that the household is not efficient, and a specific individual is more likely to invest in children if the returns are more likely to directly accrue to her.

If the household does not bargain efficiently, the consequences extend beyond investment in human capital to all investment decisions. In a Pareto efficient household, production and consumption decisions are separable: The household should choose inputs and investment levels to maximize production, and then bargain over the division of the surplus. This property will be violated if individuals make investment decisions with an eye toward maximizing the share of income that directly accrues to them. Udry (1996) shows that, in Burkina Faso, after controlling for various measures of the productivity of the field (soil quality, exposure, slope, etc.), crop, year, and household fixed effects, yields on plots controlled by women are 20% smaller than yields on men's plots.[27] This does not seem to be due to the fact that women and men have different production functions. Instead, this difference is largely attributed to differences in input intensity: In particular, much less male labor and fertilizer is used on plots controlled by women than on plots controlled by males. The fertilizer result is particularly striking, since there is ample evidence that it has sharply decreasing returns to scale. Udry estimates that the households could increase production by 6% just by reallocating factors of production within the household.

Udry explains underinvestment on women's plots by their fear of being expropriated by their husband if he provides too much labor and inputs. Another reason for inefficient investment may be the fear

[27] In Burkina Faso, as in many other African countries, agricultural production is carried out simultaneously on different plots controlled by different members of the household.

of being fully taxed by family members once the investment bears fruit. Again, an efficient household would first maximize production. However, the specific claims that a household member (or a neighbor, or a member of the extended family) can make on someone's income stream may lead to inefficient investment. Consider, for example, a situation where individuals have the right to make emergency claims on the income or savings of others in their group (for example, if someone is sick and has no money to pay for the doctors, others in his extended family have an obligation to pay the doctor). Consider a savings opportunity that will increase income by a large amount in the future (for example, saving money after harvest to be able to buy fertilizer at the time of planting). If everybody could commit not to exercise their claim during the period where the income needs to be saved, the money should be saved, and the proceeds eventually distributed to those who have a claim on it, and everybody would be better off. However, if no such commitment is possible, the individual who earned the income knows that it is likely that, should he choose to save enough for fertilizer, a claim will be exercised in the period during which the money needs to be held. He is then better off spending the money right away: Even if individuals are rational and have a low discount rate, as a group they will behave as "hyperbolic discounters", who discount the immediate future relative to today more than future periods relative to each other (Laibson (1991)). The level of investment will be low in the absence of savings opportunities offering some commitment to household members.

The fact that investments are often decided within a family, rather than by a single individual, or that the proceeds of the investment will be shared among a set of people who have not necessarily supported the cost of the investment therefore greatly complicates the incentive to invest. This may, once again, explain why some potential investments with high marginal product are not taken advantage of. It is worth noting that the lack of credit and insurance in poor countries makes these problems particularly acute there. For example, the lack of credit markets means that investment decisions are taken within the families–e.g., women cannot borrow to get the optimal amount of fertilizer on their plot–and the lack of insurance plays an important role in justifying the norms on family solidarity that seem to be hindering productive investment.

4.6 Behavioral Issues

Individuals in the developing world appear not only to be credit constrained, but also to be savings constrained. Aportela (1998) shows that when the Mexican Savings Institution "Pahnal" (Patronoto del Ahorro Nacional) expanded its number of branches through post offices in poor areas and introduced new savings instruments in the 1990s, household's savings rates increased by 3% to 5% in areas where the expansion took place. The largest increase occurred for low income households.

When an individual (or his household) has time-inconsistent preferences, formal savings instruments

may increase savings rates even when they offer very low returns (even compared to holding onto cash), because they offer a commitment mechanism. Micro-credit programs may also be understood as programs helping individuals to commit to regular reimbursements. This is particularly clear for programs, like the FINCA program in Latin America, which require that their clients maintain a positive savings balance even when they borrow.[28]

Duflo and Robinson (2003) provide direct evidence that there is an unmet demand for commitment savings opportunities among Kenyan farmers, and that investment in fertilizer increases when households have access to this opportunity. In several successive seasons, they offered farmers the option to purchase a voucher for fertilizer right after harvest (when farmers are relatively well off). The vouchers could be redeemed for fertilizer at the time when it is necessary to plant it. The take up of this program was quite high: 15% of the farmers took up the program the first time it was tried with farmers who had never encountered the NGO before. Net adoption of fertilizer increased in this group. The program was then offered to some of the farmers who had participated in the pilot program mentioned above (and thus had the opportunity to test the fertilizer for themselves, and trusted the NGO), and in this group, the take up was 80%. There is also direct evidence of the difficulty for farmers to hold on to cash: In other experiments, when farmers were given a few days before they could purchase the voucher, the take up fell by more than 50%. When they were offered the option of having the fertilizer delivered at home at the time they actually needed it (and to pay for it then), none of the farmers who had initially signed up for the program had the money to pay for the fertilizer when it was delivered. Farmers were also more likely to take up the scheme when they had cash available (for example, because the researchers had purchased their maize as part of the evaluation) than when they had maize available (even though they were offered the option to sell maize). This suggests that they are more eager to commit cash than to commit maize: Maize may be easier to save than cash.

This area of research is quite recent, and wide open. Many questions need answering, and the area of applicability is wide. For example, what is the best way to increase parents' willingness to invest in deworming drugs? Why don't all parents sign the authorization form which will grant free access to deworming to their children (Miguel and Kremer (2003))? Is it a rational decision or is it procrastination? Why does the take up of the deworming drug fall so rapidly when a small cost-sharing fee is introduced (Miguel and Kremer (2003))? Understanding the psychological factors that constrain investment decisions, and the role that social norms play in disciplining individuals, but also potentially in limiting their options, is an important area for future research. Several randomized evaluations are

[28]Karlan (2003) argues that simultaneous borrowing and savings by many clients in these institutions can be explained by the value to the small business owner of the fixed repayment schedule as a discipline device. Gugerty (2000) and Anderson and Baland (2002) interpret rotating credit and savings (ROSCAs) institutions in this light.

trying to make progress in this area. They are addressing questions as diverse as: What is the role of marketing factors in the access of poor people to loans in South Africa? Do poor people take advantage of savings products with commitment options in the Philippines? What prevents people from doing a small action that would lead them to a high return? What factors (deadline, framing, etc.) make it more likely they will do it?

A defining characteristic of these projects is that they do not involve laboratory experiments: Like the research on fertilizer in Kenya, they set up real programmes which are likely to increase poor people's investment and improve welfare if they indeed deviate from perfect rationality in the way the psychological literature suggests. In order to be fruitful, this agenda will need to avoid simply transplanting to developing countries some of the insights developed by observing behaviors in rich countries. Being poor almost certainly affects the *way* people think and decide. Decisions, after all, are based not on actual returns but on what people perceive the returns to be, and these perceptions may very well be colored by their life experience. Also, when choices involve the subsistence of one's family, trade-offs are distorted in different ways than when the question is how much money one will enjoy at retirement. Pressure by extended family members or neighbors is also stronger when they are at risk of starvation. It is also plausible that decision-making is influenced by stress. What is needed is a theory of how poverty influences decision making, not only by affecting the constraints, but by changing the decision making process itself.[29] That theory can then guide a new round of empirical research, both observational and experimental.

5 Can these micro distortions explain the macroeconomic gaps?

In this long list of potentially distorting factors there are some, like government failures or credit market failures, that most people find *a priori* plausible, and others, such as intra-family inefficiencies or learning externalities, that are more contentious, and yet others, like the behavioral factors, that have not yet been widely studied. However, even where the *prima facie* evidence is the strongest, we cannot automatically conclude that the particular distortion has resulted in a *significant* loss in productivity.

To get a sense of the potential productivity loss, we return to the Indo-U.S. comparison. Taking as given the stock of capital in India and the U.S. today, *any of the multiple distortions listed above* could have affected productivity in two different ways: First, there may be across-the-board inefficiency, because everyone could have chosen the wrong technology or the wrong product mix. Second, capital may be misallocated across firms: There may be differences in productivity across firms, either because of differences in scale, or because of differences in technology or because some entrepreneurs are more

[29] See Ray (2003) for a very nice attempt to start in this direction.

skilled than others, and the distribution of capital across these firms may be sub-optimal, in the sense that the most productive firms are too small.

Here we have chosen to emphasize this latter source of inefficiency, motivated in part by the evidence, discussed above, that tells us that there are enormous differences in productivity across firms. *We take no stance on how such an inefficient allocation of capital came about, nor on why the firms do not make the right choices, either of scale or of technologies.* Lack of access to credit is, of course, a potential explanation for both, but it could be equally explained by lack of insurance, the fear of confiscation by the government, or the gap between real and perceived returns.

The goal of this section is to set up and calibrate a simple model to investigate whether the misallocation of capital across firms within a country can explain the aggregate puzzles we started from: The low output-per-worker in developing countries, given the level of capital, and the low marginal product of capital, given the output-per-worker.

We begin with a model where the misallocation only affects the scale of production, because all the firms share the same technology. Scale obviously does not matter where there are constant returns to scale, so we need to turn to a model where there are diminishing returns at the firm level.[30] We will show that, with realistic assumptions about the relative firm sizes in India and the U.S., this model cannot go very far in explaining the aggregate facts. We then turn to a model where a better technology can be purchased for a fixed cost. We show that this model, coupled with the misallocation of capital, will help generate the aggregate facts, with realistic assumptions about the distribution of firm sizes.

5.1 A Model with Diminishing Returns

• Model setup

Consider a model where there is a single technology that exhibits diminishing returns at the firm level, say, $Y = AL^{\gamma}K^{\alpha}$, with $\gamma < 1 - \alpha$. Also, we will assume that the economy has a fixed number of firms: Without that assumption, everyone will set up multiple minuscule firms, thereby eliminating the diminishing returns effect. To justify this, we make the standard assumption that the economy has a fixed number of entrepreneurs and each firm needs an entrepreneur.

Under these assumptions, every firm would invest the same amount when markets function perfectly, but when different firms are of different sizes, the marginal product would vary across the firms and efficiency would suffer. The question is whether these effects are large enough to help us explain what we see in the data.

[30]The obvious alternative—increasing returns at the firm level—will clearly not fit the basic fact that there is more than one firm in the U.S., or that the marginal product of capital is higher in India than in the U.S.

To look at this, assume that there is a population of firms indexed by i, and that firms face a limit on how much they can borrow, so that for firm i, $K \leq K(i)$. The demand for labor from a firm that invests $K(i)$, is given by $\left[\frac{A\gamma K(i)^{\alpha}}{w}\right]^{\frac{1}{1-\gamma}}$. We assume a perfect labor market, so that given the level of capital, labor is efficiently allocated across firms. Labor market clearing then requires that

$$w = A\gamma \left[\frac{\int [K(i)^{\alpha}]^{\frac{1}{1-\gamma}} dG(i)}{\overline{L}}\right]^{1-\gamma},$$

where $G(i)$ represents the distribution of i and $\overline{L}$ is labor supply per firm. Since wages are a fraction γ of output-per-worker, it follows that output-per-worker will be

$$A \left[\frac{\int [K(i)^{\alpha}]^{\frac{1}{1-\gamma}} dG(i)}{\overline{L}}\right]^{1-\gamma}.$$

Consider an economy where, for any of the reasons we outlined above, some firms have access to more capital than others. In particular, assume that in equilibrium a fraction λ of firms get to invest an amount K^1 and the rest get to invest $K^2 > K^1$.[31] This would clearly explain why the marginal product of capital varies within the same economy. We would also expect that this inefficiency in the allocation of capital would lower productivity relative to the case where capital was optimally allocated. To get at the magnitude of the efficiency loss, note that output-per-worker in this economy will be:

$$A[\frac{\lambda(K^1)^{\alpha/(1-\gamma)} + (1-\lambda)(K^2)^{\alpha/(1-\gamma)}}{\overline{L}}]^{1-\gamma}.$$

We compare this economy with another which has a TFP of A', a labor force $\overline{L}'$ and a capital stock $\overline{K}'$, which is, in contrast with the other economy, allocated optimally across firms. To say something about productivity we also need to say how many firms there are in this economy. Let us start by assuming that the number of firms is the same. Then the ratio of output-per-worker in our first economy to that in the second is:

$$(A/A')(\frac{\overline{K}}{\overline{L}}/\frac{\overline{K}'}{\overline{L}'})^{\alpha}(\overline{L}'/\overline{L})^{1-\alpha-\gamma}[(\lambda(K^1/\overline{K})^{\alpha/(1-\gamma)} + (1-\lambda)(K^2/\overline{K})^{\alpha/(1-\gamma)}]^{1-\gamma}. \quad (6)$$

We already noted that for the India-U.S. comparison, the ratio $\frac{\overline{K}}{\overline{L}}/\frac{\overline{K}'}{\overline{L}'}$ is about 1:18. The same source (the Penn World Tables) tells us that $\overline{L}/\overline{L}'$ is about 2.7. What are reasonable values of α and γ? For $1-\alpha-\gamma$, which is the share of pure profits in the economy, we assume 20%, which is what Jovanovic and Rousseau (2003) find for the U.S. This is presumably counted as capital income, so we keep $\gamma = 0.6$ and set $\alpha = 0.2$.

First consider the case where $\lambda = 1$, so that capital is efficiently allocated in both countries. Then the productivity ratio ought to be $(A/A')(\frac{\overline{K}}{\overline{L}}/\frac{\overline{K}'}{\overline{L}'})^{\alpha}(\overline{L}'/\overline{L})^{1-\alpha-\gamma}$: Assuming that $2A = A'$, as before, because

[31] Since all firms face the same technology and there are diminishing returns to scale, this would not happen in the absence of these imperfections (all the firms should invest the same amount, $\lambda K^1 + (1-\lambda)K^2 = \overline{K}$).

of the human capital differences across these economies, the ratio works out to be $\frac{1}{2}\left(\frac{1}{18}\right)^{0.2}(\frac{1}{2.7})^{0.2} = 23\%$. Recall from equation 5 that the model with constant returns predicted that output should be 6.35 times higher in the U.S., or, equivalently that output-per-capita in India should be 15.7% the U.S. level. The 23% predicted by the current model is, of course, even further from the 9% we find in the data. The reason why this model does worse is because the production function is more concave: The concavity penalizes the U.S., which has more capital relative to India.

• Misallocation of capital: Effects on the average marginal product of capital

To bring in the effects of misallocating capital, we need to determine the size of the gap between K^2 and K^1 that we can reasonably assume. One way to calibrate these numbers is to make use of the estimate from Banerjee and Duflo (2003a) that in India there are firms where the marginal product of capital seems to be close to 100%. On the other hand, some seem to have access to capital at 9% or so, and therefore may well have a marginal product reasonably close to 9% (Timberg and Aiyar (1984)). From the production function, we know that if we assume K^1 corresponds to the firm with a marginal product of 100%, while K^2 is the firm with the marginal product of 9%, then $(K^2/K^1)^{\frac{\alpha}{1-\gamma}-1} = \frac{9}{100}$ or $K^2/K^1 = (\frac{100}{9})^2 = 123$. We can now evaluate the ratio of output-per-worker in the two economies for any given value of λ, the fraction for firms with capital stock K^1. To pin down λ, we use the fact that the average of the marginal product in India seems to be somewhere in the range of 22%. In our model, under the assumption that the marginal dollar is allocated between small firms and large firms in the same proportion as the average dollar, the average marginal product of capital is given by:

$$\frac{\lambda}{\lambda + 123(1-\lambda)}100 + \frac{(1-\lambda)123}{\lambda + 123(1-\lambda)}9.$$

Since this is equal to 22% we have that $\lambda = 0.95$. We can now compute the extent of productivity loss due to the misallocation. From equation 6, this is given by the expression $[(\lambda(K^1/\overline{K})^{\alpha/(1-\gamma)} + (1-\lambda)(K^2/\overline{K})^{\alpha/(1-\gamma)}]^{1-\gamma}$. Under the assumed values, it is approximately 0.8. In other words, the misallocation brings the productivity ratio we expect to see between India and the U.S. down from 23% to about 18%.

Relative to the neo-classical model we started from (which generates an output-per-worker in India of 15.7% of the U.S. level), moving to this model therefore does not help close the productivity gap between India and the U.S. The problem is, once again, that the additional productivity gap that the misallocation generates is more than compensated for by the effect of making the production function concave while keeping the number of firms fixed.

What does this model predict for the marginal product of capital in the U.S? Since $K^2 = 123K^1$, $\overline{K_I} = [0.955 + 123(0.045)]K^1 = 6.5K^1$. Therefore, $K^2 \approx 19\overline{K_I}$. Now since $\left(\frac{\overline{K}}{\overline{L}}\right)_I / \left(\frac{\overline{K}}{\overline{L}}\right)_U$ is about 1/18

and $\overline{L}_I/\overline{L}_U$ is about 2.7, $\overline{K}_I/\overline{K}_U = 0.15$. Therefore $K^2/\overline{K}_U = 2.85$. The ratio of the marginal product of capital in the large Indian firms to that in the average U.S. firm under the assumption that TFP is twice as high in the U.S. (because workers in India are about 30% as productive as workers in the U.S.), is given by the expression

$$\begin{aligned}(\frac{A^I}{A^U})^{\frac{1}{1-\gamma}}(K^2/\overline{K}_U)^{\frac{\alpha}{1-\gamma}-1}(w_U/w_I)^{\frac{\gamma}{1-\gamma}} &= (\frac{A^I}{A^U})^{\frac{1}{1-\gamma}}(K^2/\overline{K}_U)^{\frac{\alpha}{1-\gamma}-1}(y_U/y_I)^{\frac{\gamma}{1-\gamma}} \\ &= (\frac{1}{2})^{\frac{5}{2}}(2.85)^{-1/2}(11)^{3/2} \approx 3.8.\end{aligned}$$

This predicts return on capital in the U.S. to be a quarter of the 9% return we assumed for the large firms in India, which is clearly much too low (the U.S. rate is usually estimated to be 9%).

One way to resolve both these problems is to give up the assumption that the two economies have same number of firms. Suppose the U.S. had $\lambda > 1$ times as many firms as India: Then the labor productivity ratio computed above would have to be divided by $\lambda^{1-\alpha-\gamma}$. If λ were equal to 32, the ratio of labor productivity in India to that in the U.S. would be 9%, which is what we find in the data.

Of course, increasing the number of firms in the U.S. will tend to make the average firm in the U.S. smaller: Even with the same number of firms in the two countries, the fact that the biggest firms in India have about 18 times the average capital stock means that they are about 3 times the size of U.S. firms, which seems implausible. If there are 32 times as many firms in the U.S., the average U.S. firm would be about a 1/100th of the biggest Indian firm, close to 25% in the Indian size distribution. This seems entirely counterfactual.

• Predictions on the distribution of marginal product of capital within countries

We see another problem with this model when we focus on the comparison of marginal products within countries—this is not something that can be fixed by manipulating the number of firms. Table 1 lists, for nine of the largest industries in India (where industry is defined at 3 digit level) outside of agriculture, known for having a substantial presence of small enterprises, some measures of variation in firm sizes (where size is measured by the net fixed capital in year 2000). We see that the ratio of the 95th percentile firm to the 5th percentile firm in the median industry is approximately 1,600:1.[32] Given the production function, we know that the marginal return on capital in the two firms should differ by a factor of $1600^{1/2} = 40 : 1$. Since the biggest firms pay about 9% for their capital, the smaller firms must have a marginal product that exceeds 360%, which seems implausible.

Finally, this particular parameterization of the model assumes an industry structure that is rather extreme. In the industry described by our model, the large firm in our model is 123 times the size of the small firm. In the ASI data, even the 95th percentile firm in the median industry is no more than 72

[32]The median industry is the Textile Garment Manufacturing industry.

Table 1: Distribution of firm size (Annual Survey of Industry, 2000)

	95-5 ratio	median-5 ratio	mean-5 ratio	5th percentile
Manufacture of Pasteurized Milk	1007	95	216	61466
Flour milling	786	150	285	29899
Rice milling	1392	90	620	5681
Cotton Spinning	22300	440	5423	12870
Cotton Weaving	3093	31	1292	14159
Textile garment manufacture	1581	104	410	22461
Curing raw hides and skins	235	10	53	37075
Manufacture of footwear	2639	122	683	21825
Manufacture of car parts	1700	29	504	84103

times the 25th percentile firm. The firm that is $\frac{1}{123}$ times the 95th percentile firm in the median industry is around the 20th percentile in the size distribution. More than 50% of the capital stock in the Indian economy is in firms that are bigger than the "small" firm and smaller than the "large" firm as we defined them here. If we tried to use a more realistic distribution of firm sizes, it would make it even harder to explain the productivity gap between India and the U.S.: Moving weight closer to the mean would dampen the effect of concavity that is at the heart of our theory.

• Taking Stock

To sum up, moving to this more sophisticated model does not help us fit the macro facts better. It obviously does suggest a simple theory of the cross-sectional variation in returns to capital, which is entirely absent from the model with constant returns, at the cost of predicting an implausibly high degree of variation in firm sizes. Moreover, it only helps to explain the productivity gap between India and the U.S. if we assume that the biggest firms in India are almost six times the average U.S. firms in the same sector.

The next section introduces an alternative model where firms differ both in scale and in technology, but still retains the assumption that there is no inherent difference between these alternative investors.

5.2 A Model with Fixed Costs

• Model Setup

Consider a world where setting up requires a fixed start-up cost in addition to an entrepreneur, but once these are in place, capital and labor get combined as in a standard Cobb-Douglas with diminishing

returns. This fixed cost could come from many sources: Machines come in certain discrete sizes and even the smallest machine may be expensive from someone's point of view. Buildings, likewise, are somewhat indivisible, at least by the time we come down to a single room or less. Marketing and building a reputation may also require an indivisible up-front investment—Banerjee and Duflo (2000) describe the costs that a new firm in the customized software industry has to pay in terms of harsh contractual terms, until it has a secured reputation. Turning to investment in human capital, it also appears that the first five years or so of education may have much lower returns than the next few years, which in effect makes the first few years of education a fixed cost.[33] Finally, as emphasized by Banerjee (2003a) the fixed cost may be in the financial contracting that the firm has to go through—starting loans are often expensive because the lender cannot trust the borrower with a big loan and when the loan is small, the fixed costs of setting up the contract loom large.

Formally, we assume a production function $y = A(K - \underline{K})^{\alpha} L^{\gamma}$. Since we continue to assume that the firm can buy as much labor as it wants, the production function can be rewritten as:

$$A^{\frac{1}{1-\gamma}} \left[\frac{\gamma}{w}\right]^{\frac{\gamma}{1-\gamma}} [K - \underline{K}]^{\frac{\alpha}{1-\gamma}}. \tag{7}$$

We continue to assume that $\gamma + \alpha < 1$, so that there are diminishing returns. The average cost function in this world has the classic Marshallian shape: Average costs go down first as the fixed costs get amortized over more and more output and then start to rise again. The optimal scale of production is given by the equality of the marginal and average product of capital, which reduces to:

$$K = \underline{K} \frac{1-\gamma}{1-\gamma-\alpha}.$$

We allow firms the option of choosing between alternative technologies. Assume that there are three alternative technologies available, characterized by three different levels of the fixed cost, $\underline{K}_1$, $\underline{K}_2$ and $\underline{K}_3$, three differing levels of labor and capital intensity, $\{(\alpha_1, \gamma_1), (\alpha_2, \gamma_2), (\alpha_3, \gamma_3)\}$ and three correspondingly different levels of productivity, A_1, A_2 and A_3. We make the usual assumption that a higher cost buys a higher levels of TFP, i.e., that $\underline{K}_1 \leq \underline{K}_2 \leq \underline{K}_3$ and $A_1 \leq A_2 \leq A_3$.

Compared to a Cobb-Douglas model with diminishing returns, this formulation has a number of advantages. First, it allows firms to have large differences in size without necessarily large differences in the marginal product of capital, since they could be using different technologies. The fact that there are firms in the same industry operating at very different scales posed a problem for the model with diminishing returns because the implied variation in the marginal product of capital seems implausibly large. Second, the fact that production requires a fixed cost helps explain why, despite the diminishing

[33] All the estimates (14) we could find of Mincerian returns at different levels of education suggest that in developing countries the marginal benefits of a year of education increase with the level of education (in the U.S., it appears to be very flat). Schultz (2001) finds the same result in his study of six African countries.

returns from technology, we do not see people setting up a very large number of very small firms, thereby completely eliminating the diminishing returns effect. In this case, we can let the number of firms be determined by what people are willing to invest, in combination with what we know about the fixed costs (actually as noted below, we cheat slightly on this point, but only because it simplifies the calculations). Third, because we allow the number of firms to be determined endogenously, there are fewer overall diminishing returns when we compare the U.S. and India, which helps explain why the productivity gap is so large and why interest rates are not lower in the U.S. Fourth, as noted above, this model generates a unique optimal scale of production, which would provide a reason why the most productive Indian and U.S. firms would look relatively similar. Finally, making this assumption alters the nature of the link between the marginal product of capital and its average product. With a Cobb-Douglas, the ratio of the average product is always proportional to the marginal product. Here, the average product starts lower than the marginal product but grows faster and eventually becomes larger than the marginal product. In other words, as firm size goes up the ratio of the marginal product of capital to its average product goes down, at least initially. This would suggest that the ratio of the average products of capital in India and the U.S. should be less than the ratio of the marginal products, and indeed we find that while output-per-worker is 11 times larger in the U.S., capital-per-worker is 18 times as large, implying an average product ratio of about 1.6:1, as against the 2.5:1 ratio of marginal product delivered by the standard Cobb-Douglas model. This is clearly an *a priori* advantage of this formulation, since, as we noted in section 3, the proportionality between the average product and the marginal product prevents any model based on a Cobb-Douglas production function to fit these facts.

In order to impose restrictions on the parameters of the model, we make use of the industry data described in table 1. We describe the representative Indian industry by a 3-point distribution of firm sizes, with fractions λ_1, λ_2, and λ_3 at K_1, K_2 and K_3. The first group of firms is made up from the bottom 10% of the distribution of firms, and we assigned to them the size of the firm at the 5th percentile of the actual size distribution in the data. Likewise, we assume that the top 10% of all firms are in the group of "large firm", and that their size is that of the firm at the 95th percentile of the firm size distribution.[34] The rest we assign to the middle category, whose size we set at the mean for the distribution. We assume that the largest firm is 1,600 times as big as the smallest firm, which is roughly the median value of these ratios across these nine industries in our data.

These parameter values imply that the mean firm size in the industry will be 800 times as large as the 5th percentile firm, which is higher than the mean in the median industry in our data (500 times), but well within the existing range in the data. Once again we are interested in the within-economy variation

[34] We pick the 5th and the 95th percentile to make the difference in the returns to capital between the biggest and smallest firms as large as possible.

in returns to capital. We therefore assume, as before, that the small firms have a marginal product of 100% while the medium-sized firms have a marginal product of just 9%.

The more unorthodox assumption is that the large firms also have a marginal product of 100%. While clearly somewhat artificial, this is meant to capture the idea that the best technology is expensive and only the biggest firms in India can afford to be at the cutting edge, an idea that is very much in the spirit of the McKinsey Global Institute's study of a number of specific Indian industries. However, they are still relatively small and therefore the marginal returns on an extra dollar of investment are very high. The rest of the firms use cheaper (i.e., lower $\underline{K}$) but less effective technologies. In particular, the small firms are simply too small (which explains their high marginal product), and the middle category consists of firms that have exhausted the potential of the mediocre technology that they can afford but are too small to make use of the ideal technology.

How plausible is our assumption about industry structure? The average capital stock of the 95th percentile firms in the median industry was Rs. 36 million, which puts them at a size just above the category of firms that are the focus of Banerjee and Duflo (2003a). The point of that paper was that a subset of these firms (the firms that attracted the extra credit after the policy change) had marginal returns on capital of 100% or more. Therefore, it is not absurd to assume that the large firms in our model economy have very high returns. Once we accept the idea that some large firms are very productive, given that the average marginal product is probably close to 22%, it is obviously very likely that there are many smaller firms that have a *lower* marginal product than the largest firms. Indeed, when we calculate the average marginal product based on our assumptions, under the premise that the marginal dollar is distributed across the three size categories in the ratio of their share in the capital stock, the average marginal product turns out to be about 27%.

Even with this long list of assumptions, we do not have enough information to compute output per worker in our model economy—there are several remaining degrees of freedom. First, we need to choose units: Our assumption, which simplifies calculations, is that capital is measured in multiples of the small firm. Finally, we assume that $\underline{K}_1 = 0$, $\underline{K}_2 = 100$, and $\underline{K}_3 = 800$. The assumption that $\underline{K}_3 = 800$, implies that the biggest Indian firms (which have 1,600 units of capital) are operating at the bottom of the average cost curve—given by $\underline{K}\frac{1-\gamma}{1-\gamma-\alpha}$.[35]

• Results: Output-per-worker and average marginal product of capital

Under these assumptions, we can use the assumed marginal products to solve for A_1, A_2 and A_3.

[35]This is where we cheat, since with decreasing returns to scale, there could again be an infinity of very small firms, so that all the firms should be in the small group. We can prevent this if we assume that the smallest feasible firm size is actually ϵ greater than zero, and only a certain number of entrepreneurs are able (or willing) to invest at least ϵ.

According to these calculations, TFP in the medium firms is about 1.4 times bigger than that in the small firms, and that in the large firms is about 2.7 times that in the medium firms. Nevertheless, given the assumed limits on how much they can invest, each category of firms is optimizing by choosing its current technology. However, large gains in productivity are obviously possible if the economy can reallocate its capital so that all the firms adopt the most productive technology.

To see how large this gain may be, we do another India-U.S. comparison. Once again we assume that the U.S. takes full advantage of the available technology. In other words, every firm in the U.S. operates the best technology at the optimal scale, i.e., each of these firms operates technology 3 and has 1,600 units of capital. The distribution of firm sizes in India, by contrast, includes a large fraction of firms that neither use the best technology nor operate at the optimal scale. The implicit assumption is that in the U.S. there are enough people who are able and willing to invest 1,600 units if there is any money to be made, but this is not true in India because of borrowing constraints or other reasons.

A series of straightforward calculations gives the expression for the ratio of output-per-worker, which is also the ratio of wages in the two economies:

$$
\begin{aligned}
& (\frac{y_I}{y_U})^{\frac{1}{1-\gamma}} = (\frac{w_I}{w_U})^{\frac{1}{1-\gamma}} \\
= \quad & \frac{N_I}{N_U}\frac{L_U}{L_I} \times \\
& \frac{[\lambda_1(A_I)^{\frac{1}{1-\gamma}}(K_1-\underline{K_1})^{\frac{\alpha}{1-\gamma}} + \lambda_2(A_IA_2/A_1)^{\frac{1}{1-\gamma}}(K_2-\underline{K_2})^{\frac{\alpha}{1-\gamma}} + \lambda_3(A_IA_3/A_1)^{\frac{1}{1-\gamma}}(K_3-\underline{K_3})^{\frac{\alpha}{1-\gamma}}]}{(A_UA_3/A_1)^{\frac{1}{1-\gamma}}(K_3-\underline{K_3})^{\frac{\alpha}{1-\gamma}}},
\end{aligned}
$$

where N_I and N_U are the numbers of firms in India and the U.S., and A_I and A_U represent the base levels of TFP. The only reason that $A_I \neq A_U$ is, as before, that the human capital levels vary. We continue to assume that $A_U = 2A_I$. N_I/N_U can be computed from the fact that the total demand for capital from these firms must exhaust the supply of capital: i.e.,

$$\frac{K_I}{K_U} = \frac{N_I[\lambda_1K_1 + \lambda_2K_2 + \lambda_3K_3]}{N_UK_3},$$

which, given the assumed parameter values, implies that $N_I/N_U = 0.3$, which can then be used to calculate $\frac{y_I}{y_U}$, which turns out to be almost exactly 1/10, not too far from the 1/11 that we found in the data.

We can also derive, as before, what this model tells us about the marginal product of capital in the U.S. Using the expression derived in the previous sub-section, it is easily shown that the ratio of the marginal product of capital in the U.S. to that in the biggest and best Indian firms will be given by $(\frac{A^I}{A^U})^{\frac{1}{1-\gamma}}(w_U/w_I)^{\frac{\gamma}{1-\gamma}}$,[36] which turns out to be 6.45. Given that the biggest Indian firms have a marginal

[36]The fact that the biggest firms in India are the same size as any U.S. firm obviously simplifies the calculation.

product of 100%, the average U.S. firms should have a marginal product of 100/6.45=15.5%. This is obviously higher than the average stock market return, but hardly beyond the reasonable range.

• Distribution of firm sizes

The most obvious advantage of the fixed cost approach is that we do not obtain the unreasonably large gap in the marginal products of capital between small and large firms within the same economy, which came out of the previous model. This underscores the importance of using evidence on cross-sectional differences within an economy to assess the validity of alternative models.

Finally, the success of this model in explaining the productivity gap depends, as in the case of the previous model, heavily on the assumption that the U.S. has many more firms than India. However, while in that model we needed the U.S. to have 32 times as many firms as in India to fit the observed productivity gap, here we are doing very well with the U.S. having $3\frac{1}{3}$ times as many.

How reasonable is the assumption that the U.S. has more firms than India? This is not an easy question to answer, mainly because we have no clear sense of what should count as a firm: Both economies have enormous numbers of tiny firms that reflect what people do on the side. In India these "firms" are concentrated in a few sectors, such as retailing or the collection of leaves, wood or waste products, which require little or no skills and can be done on part-time basis. In the U.S., the equivalent would be the numerous ways in which you end up owning a small business for tax purposes, such as part-time consulting, renting out part of your home, part-time telemarketing, etc. It is not clear which of these should count as legitimate firms from the point of view of our model and which of these should not.

A way to restate the same point is that by focusing on the median industry in the ASI data we have effectively ignored the industries (like the ones listed above) which attract all those in India who have nowhere better to go. While there are only a few such industries, they are enormous, and quite unlike the rest of the industries: Among the industries listed in the table above, cotton spinning is probably most like what one of these industries looks like, and it is apparent that it is quite different from the rest—there are many more tiny firms.

Adopting a model of the industry structure in India that has more small and inefficient firms, and therefore less large and efficient firms, is in many ways very much like assuming that there are fewer firms in India. It is easy to show that if we re-parameterize the model in this section to reduce the fraction of large firms (firms with 1,600 units of capital) to 3% (from 10%), but assume that the two economies have the same number of firms, output-per-worker in India would once again be 10% of what it is in the U.S.

• Why doesn't capital flow to India?

Finally we subject this model to an additional test: The fact that in our model there are firms in India

with returns in the neighborhood of 100% would suggest that there are many unexploited opportunities. We have already argued that there are many reasons why a U.S. bank could not just lend to an Indian firm, and thereby benefit from these opportunities. Nor is it easy for an American to borrow money in the U.S. and set up a firm in India: Once he is in India he may be beyond the reach of U.S. law and for that reason alone, lenders will shy away from him. What is much more plausible, however, is that a U.S. entrepreneur moves to India to invest his money in these opportunities. The question is why this does not happen more often.

There are some obvious answers to this question: If the reason why these opportunities have not already been taken advantage of is the lack of secure property rights in India, there is no reason why foreigners would be particularly keen to invest in India. On the other hand, if the problem is that Indians do not have the capital or that they fear the risk exposure or that they are simply unaware of the opportunity, to take some plausible alternatives, a well-diversified wealthy U.S. investor may well be attracted to move to India and start a firm.

How much money would such an investor make? To answer this we start by observing from (7) that the production function in the largest Indian firms can be written as $C(K-800)^{1/2}$, where $C = A_3^{\frac{1}{1-\gamma}} \left[\frac{\gamma}{w}\right]^{\frac{\gamma}{1-\gamma}}$. Of this, a fraction 3/5 goes to wages. Profits are therefore given by $\frac{2}{5}C(K-800)^{1/2}$. Since this firm has 1,600 units of capital, and the marginal product of capital in this firm was assumed to be 100%, it follows that

$$\frac{1}{5}C(800)^{-1/2} = 1,$$

or

$$C = 141.42.$$

The opportunity cost of capital for a U.S. investor is 9%. The optimal investment in this Indian firm for a U.S. investor who can invest as much as he wants will be given by the solution to

$$(141.42)(0.2)(K-800)^{-1/2} = 0.09.$$

This tell us that the optimal investment is $K = 99564$. The total after-wage income generated by the firm is $(0.4)(141.42)(99564-800)^{1/2} = 17777$. This is in units of the smallest firm. We know that the biggest firms in our model are 1,600 times as large as the smallest firms and from the table above, such firms have Rs. 36 million worth of capital in the median industry. The smallest firm therefore has Rs.22,500 worth of capital, which implies that the U.S. investor will earn 17777(22500)=Rs. 400 million on his investment of (99564)(22500)=Rs. 2.24 billion. This is a net gain of about Rs. 200 million, or about 4 million dollars.

Is this a large enough gain to tempt someone to leave his home and family and settle in India? For

someone with an average income, obviously. But no one with an average income has 50 million dollars of his own that he is willing to put into a single project in India. Anyone who is willing to do it has to be very rich indeed—he must have $50 million several times over. How many people are so wealthy that they are willing to give up their life in the U.S. for an extra $4 million per year?

In other words, while the model developed in this section generates very large productivity losses, it does not offer any one person the possibility of arbitraging these unexploited opportunities to become enormously rich. This is because diminishing returns set in quite fast.

5.2.1 Taking Stock

We started by describing some of the major puzzles left unanswered by the neo-classical model, and in particular the productivity gap between rich and poor countries. The coexistence of high and low returns to investment opportunities, together with the low average marginal product of capital, suggested that some of the answer might lie in the misallocation of capital. The microeconomic evidence indeed suggests that there are some sources of misallocation of capital, including credit constraints, institutional failures, and others. In this section, we have seen that, combined with multiple technological options and a fixed cost of upgrading to better technologies, a model based on misallocation of capital does quite well in terms of explaining the productivity gap. The value of the marginal productivity of capital in the U.S. predicted by this model is only marginally too high, and the degree of variation in the marginal product of capital within a single economy that the model requires is not implausibly large.

Of course the model does make unrealistic assumptions—there is, for example, surely some amount of inefficiency in the U.S., and some U.S. firms are surely more productive than others. On the other hand, we have also ignored many reasons why Indian firms may be less efficient than they are in our model. For example, our current model assumes that only 10% of the firms, who use less than 1% of the capital stock and produce less than 1% of the output, use the least efficient technology whereas the MGI report on the apparel sector tells us that almost 55% of the output of the sector is produced by tailors who still use primitive technology. We also assumed that 10% of Indian firms are as productive as the best U.S. firms. Clearly that fraction could be smaller.

We also assumed that everyone is equally competent. In the real world, imperfect credit markets, for example, drives down the opportunity cost of capital and this encourages incompetent producers to stay in business. In the model, we assume that all large firms earn high returns but in reality there are probably some large firms that have much lower productivity (anywhere down to 9% per year would be consistent with our model). This too will drive down productivity. In a recent paper, Caselli and Gennaioli (2002) try to calibrate the impact of this factor in the context of a dynamic model with credit constraints. They show that in steady state this can generate productivity losses of 20% or so. We will

argue in the next section that this severely understates the potential productivity gap starting from an arbitrary allocation of capital.

6 Towards a Non-aggregative Growth Theory

6.1 An Illustration

The presumption of neo-classical growth theory was that being a citizen of a poor country gives one access to many exciting investment opportunities, which eventually lead on to convergence. The point of the previous section was to argue that most citizens of poor countries are not in a position to enjoy most of these opportunities, either because markets do not do what they ought to or the government does what it ought not to, or because people find it psychologically difficult to do what is expected of them.

What can we say about the long-run evolution of an economy where there are rewarding opportunities that are not necessarily exploited? In this section we will explore this question under the assumption that the only source of inefficiency in this economy comes from limited access to credit. The goal is to illustrate what non-aggregative growth theory might look like, rather than to suggest an alternative canonical model.

The model we have in mind is as follows: There are individual production functions associated with every participant in this economy that are assumed to be identical and a function of capital alone ($F(K)$) but otherwise quite general. In particular, we do assume that they are concave. Individuals maximize an intertemporal utility function of the form:

$$\begin{aligned} \sum_{t=0}^{\infty} \delta^t U(C_t), 0 &< \delta < 1 \\ U(C_t) &= \frac{c^{1-\phi}}{1-\phi}, \phi > 0. \end{aligned}$$

People are forward-looking and at each point of time they choose consumption and savings to maximize lifetime utility. However, the maximum amount they can borrow is linear and increasing in their wealth and decreasing in the current interest rate: An individual with wealth w can borrow up to $\lambda(r_t)w$. Credit comes from other members of the same economy and the interest rate clears the credit market. We do not assume that everyone starts with the same wealth, but rather that at each point of time there is a distribution of wealth that evolves over time.

This model is a straightforward generalization of the standard growth model. What it tells us about the evolution of the income distribution and efficiency depends, not surprisingly, on the shape of the production function.

The simplest case is that of constant returns in production. In this case, inequality remains unchanged over time, and production and investment is always efficient.

With diminishing returns, greater inequality can lead to less investment and less growth, because the production function is concave. However, inequality falls over time and in the long run no one is credit constrained, although we do not necessarily get full wealth convergence. The long run interest rate converges to its first best level, and hence investment is efficient. To see why this must be the case, note first that because of diminishing returns the poor always have more to gain from borrowing and investing than the rich. In other words, the rich must be lending to the poor. As long as the poor are credit constrained, they will earn higher returns on the marginal dollar than their lenders, i.e., the rich (that is what it means to be credit constrained). As a result, they will accumulate wealth faster than the rich and we will see convergence. This process will only stop when the poor are no longer credit constrained, i.e., they are rich enough to be able to invest as much as they want.

With increasing returns, inequality increases over time; we converge to a Gini coefficient of 1. Wealth becomes more and more concentrated with only the richest borrowing and investing. Because there are increasing returns, this is also the first best outcome. The logic of this result is very similar to the previous one: Now it is the rich who will be borrowing and the poor who will be lending, with the implication that the rich are the ones who are credit constrained and the ones earning high marginal returns. Therefore, they will accumulate wealth faster and wealth becomes increasingly concentrated.

Finally we consider the case of "S-shaped" production functions, which are production functions that are initially convex and then concave. The Cobb-Douglas with an initial set-up cost discussed at length in section 5.2 is a special case of this kind of technology.

What happens in the long run in this model depends on the initial distribution of income. When the distribution is such that most people in the economy can afford to invest in the concave part of the production function, the economy converges to a situation that is isomorphic to the diminishing returns case, with the entire population "escaping" the convex region of the production function.

The more unusual case is the one where some people start too poor to invest in the concave region of the production function. The poorer among such people will earn very low returns if they were to invest and therefore will prefer to be lenders. Now, as long as the interest rate on savings is less than $1/\delta$, they will decumulate capital (since the interest is less than the discount factor) and eventually their wealth will go to zero. On the other hand, anyone in this economy who started rich enough to want to borrow will stay rich, even though they are also dissaving, in part because at the same time they benefit from the low interest rates. The economy will converge to a steady state where the interest rate is $1/\delta$, those who started rich continue to be rich and those who started poor remain poor (in fact have zero wealth).

This is classic poverty trap: Moreover, since no one escapes from poverty, nor falls into it, there is a

continuum of such poverty traps in this model. This kind of multiplicity is, however, fragile with respect to the introduction of random shocks that allow some of the poor to escape poverty and impoverish some of the rich.

Even in a world with such shocks there can be more than one steady state: The reason is that the presence of lots of poor people drives down interest rates, and low interest rates make it harder for the poor to save up to escape poverty even with the help of a positive shock. As a result, in an economy that starts with lots of poor people, a greater fraction of people may remain poor.

The key to this multiplicity is the endogeneity of the interest rate. It is the pecuniary externality that the poor inflict on other poor people that sustains it. This is why such poverty traps are sometimes called *collective poverty traps*, in contrast to the *individual poverty traps* described above.

The investigation of the evolution of income distribution in models with credit constraints and endogenous interest rates goes back to Aghion and Bolton (1997). Matsuyama (2000, 2003) and Piketty (1997) emphasize the potential for collective poverty traps in a variant of this model, without the forward-looking savings decisions.

This class of models is a part of a broader group of models which study the simultaneous evolution of the occupational structure, factor prices and the wealth distribution in a model with credit constraints. Loury (1981) studied this class of models and showed that in the long run the neo-classical predictions tend to hold as long as the production function is concave. Dasgupta and Ray (1986) and Galor and Zeira (1993) provide examples of individual poverty traps in the presence of credit constraints and S-shaped production functions. Banerjee and Newman (1993) show the possibility of a collective poverty trap in a model with a S-shaped production function which is driven by the endogeneity of the wage–essentially high wages allow workers to become entrepreneurs easily, which keeps the demand for labor, and hence wages, high. Recent work by Buera (2003) shows that the multiplicity results in Banerjee and Newman survive in an environment where savings is based on expectations of future returns.[37] Ghatak, Morelli and Sjostrom (2001, 2002) and Mookerjee and Ray (2002, 2003) explore related but slightly different sources of individual and collective poverty traps.

6.2 Can we take this model to the data?

Models like the one we just developed (as well as political economy models that we do not discuss here[38]) have been invoked as motivation for a large empirical literature on the relationship between inequality

[37] On the possibility of collective poverty traps, see also Lloyd-Ellis and Bernhardt (2000), and Mookerjee and Ray (2002, 2003).

[38] See Alesina and Rodrik (1994), Persson and Tabellini (1991) and Benhabib and Rustichini (1998). For a contrarian point of view, arguing that the premise of the political economy model argument does not hold true in the data, see Benabou (1996).

and growth in cross-country data. In 1996, Benabou cited 16 studies on the question, and the number has been growing rapidly since then, in part due to the availability of more complete data sets, due to the effort of Deininger and Squire (see Deininger and Squire (1996)), expanded by the World Institute for Development Economics Research (WIDER). However, it is not clear that if we were to take this class of models seriously, they would justify estimating relationships like the ones that are in the literature: First because the exact form of the predicted relationship between inequality and growth depends on the shape of the production function. Imposing the assumption that there are diminishing returns helps in this respect, but with this assumption functional form issues loom large. Finally, it is not clear how, given the model's structure, we can avoid running into serious identification problems.

In this section, we evaluate whether, given these concerns, estimating the relationship between inequality and growth in a cross-country data set remains useful. Having concluded that it has, at best, very limited use, we discuss an alternative approach based on calibrating non-aggregative models using micro data.

6.2.1 What are the empirical implications of the above model?

Functional Form Issues With constant returns to scale, distribution is irrelevant for growth. With diminishing returns, an exogenous mean-preserving spread in the wealth distribution in this economy will reduce future wealth and, by implication, the growth rate. However, the impact depends on the level of wealth in the economy: Once the economy is rich enough that everyone can afford the optimal level of investment, inequality should not matter. The estimated relationship between inequality and growth should therefore allow for an interaction term between inequality and mean income. Moreover, an economy closer to the steady state has both lower inequality and lower growth. This has two implications for the estimation of the inequality growth relationship. First, the fact that the economy becomes more equal as it grows tends to generate a spurious positive relation between growth and inequality, both in the cross-section as well as in time-series. As a result, both the cross-sectional and the first differenced (or fixed effects) estimates of the effect of inequality on growth run the risk of being biased upwards, compared to the true negative relation that we might have found if we had compared economies at the same mean wealth levels. Moreover, consider a variant of the model where there are occasional shocks that increase inequality. Since the natural tendency of the economy is towards convergence, we should expect to see two types of changes in inequality: Exogenous shocks that increase inequality and therefore reduce growth and endogenous reductions in inequality that are also associated with a fall in the growth rate. In other words, measured changes in inequality in either direction will be associated with a fall in growth.

Controlling properly for the effect of mean wealth (or mean income), is therefore vital for getting

meaningful results. The usual procedure is to control linearly (as in most other growth regressions) for the mean income level at the beginning of the period. It is, however, not clear that there is any good reason why the true effect should be linear. Moreover, it seems plausible that different economies will typically have different λs, and therefore will converge at different rates.

The model also tells us that while initial distribution matters for the growth rate, it only matters in the short run. Over a long enough period, two economies starting at the same mean wealth level will exhibit the same average growth rate. In other words, the length of the time period over which growth is measured will affect the strength of the relationship between inequality and growth.

The preceding discussion assumed that the interest rates converged. As we noted, that does not need to be the case. If we do not assume it, variants of the simple concave economy may no longer converge, even in the weaker sense of the long-run mean wealth being independent of the initial distribution of wealth. Intuitively, poor economies will tend to have high interest rates, and this in turn will make capital accumulation difficult (note that $\lambda' < 0$) and tend to keep the economy poor.[39] This effect reinforces the claim made above that inequality matters most in the poorest economies.[40] This economy can have a number of distinct steady states that are each locally isolated. This means that small changes in inequality can cause the economy to move towards a different and further away steady state, making it more likely that the relationship will be non-linear.

With increasing returns, growth rates increase with a mean preserving spread in income. As the economy grows, it also becomes more unequal. Interpreting the relationship between inequality and growth is difficult even after controlling for convergence.

In the S-shaped returns case, the relationship between inequality and growth can be negative or positive depending on the initial distribution, and the size of the increase. For example, if everybody is very poor (on the left of the convex zone), a small increase in inequality will reduce growth, but increasing inequality enough may push more people to the point where they are able to take advantage of the more efficient technology, and increases in inequality will increase growth. The relation between inequality and growth delivered by this model is clearly non-monotonic. Moreover, the strong convergence property does not hold in general. In other words, the growth rate of wealth may *jump up* once the economy is rich enough, with the obvious implication that economies with higher mean wealth will not necessarily grow more slowly. In other words, the effect of mean wealth, that is the so-called convergence effect, may not be monotonic in this economy. Linearly controlling for mean wealth therefore does not guarantee that

[39] See Piketty (1997). For a more general discussion of the issue of convergence in this class of models, see Banerjee and Newman (1993).

[40] There is, however, a counteracting effect: Poorer economies with high levels of inequality may actually have low interest rates because a few people may own more wealth than they can invest in their own firms, and the rest may be too poor to borrow. For a model where this effect plays an important role, see Aghion and Bolton (1997).

we will get the correct estimate of the effect of inequality. It is worth noting that this economy will have a connected continuum of steady states. This means that after a shock the economy will not typically return to the same steady state. However, since it does converge to a nearby steady state, this is not an additional source of non-linearity.

Identification Issues Even if we could agree on a specification that is worth estimating, it is not clear how we can use cross-country data to estimate it. Countries, like individuals, are different from each other. Even in a world of perfect capital markets, countries can have very different distributions of wealth because, for example, they have different distributions of ability. There is no causal effect of inequality on growth in this case, but they could be correlated for other reasons. For example, cultural structures (such as a caste system) may restrict occupational choices and therefore may not allow individuals to make proper use of their talents, causing both higher inequality and lower growth. Conversely, if countries use technologies that are differently intensive in skilled labor, those countries using the more skill intensive technology can have both more inequality and faster growth.

As we discussed in detail above, countries have different kinds of financial institutions, implying differences in the λ's in our model. Our basic model would predict that the country with the better capital markets is likely both to be more equal and to grow faster (at least once we control for the mean level of income). The correlation between inequality and growth will therefore be a downwards-biased estimate of the causal parameter, if the quality of financial institutions differs across countries.[41]

If these country specific effects were additive, one could control for them by including a country fixed-effect in the estimated relationship (or by estimating the model in first difference). This strategy will be valid only under the assumption that changes in inequality are unrelated to unobservable country characteristics that are correlated with changes in the growth rate. While this is a convenient assumption, it has no reason to hold in general. For example, skill-biased technological progress will lead both to a change in inequality and a change in growth rates, causing a spurious positive correlation between the two. To make matters worse, we have to recognize the fact that λ itself (and therefore the effect of inequality on growth at a given point in time) may be varying over time as a result of monetary policies or financial development, and may itself be endogenous to the growth process.[42]

The more general point that comes out of the discussion above is that unless we assume capital

[41] Allowing λ to vary also implies that the causal effects of inequality will vary with financial development (which is how Barro (2000) explains his results). The OLS coefficient is therefore a weighted average of different parameters, where the weights are the country-specific contributions to the overall variance in inequality (Krueger and Lindahl (1999)). It is not at all clear that we are particularly interested in this set of weights.

[42] See Acemoglu and Zilibotti (1994), and Greenwood and Jovanovic (1999), for theories of growth with endogenous financial development.

markets are extremely efficient (which, in any case, removes one of the important sources of the effect of inequality), changes in inequality will be partly endogenous and related to country characteristics which are themselves related to changes in the growth rate. Identifying the effect of inequality by including a country fixed-effect would not necessarily solve all the endogeneity problems. Moreover, as we discussed above, the theory suggests that the specification should allow for non-linear functional forms, and interaction effects, which will be difficult to accommodate with a fixed effect specification.

6.2.2 Empirical Evidence

The preceding discussion suggests that empirical exercises using aggregate, cross-country data to estimate the impact of inequality and growth will be extremely difficult to interpret. The results are also likely to be sensitive to the choice of specification. This may explain the variety of results present in the literature. A long literature (see Benabou (1996) for a survey) estimated a long run equation, with growth between 1990 and 1960 (say) regressed on income in 1960, a set of control variables, and inequality in 1960. Estimating these equations tended to generate negative coefficients for inequality. As the discussion in the previous subsection suggests, there are many reasons to think that this relationship may be biased upward or downwards. To address this problem, Li and Zou (1998) and Forbes (2000) used the Deininger and Squire data set to focus on the impact of inequality on short run (5 years) growth, and introduced a linear fixed effect.[43] The results change rather dramatically: The coefficient of inequality in this specification is positive, and significant. Finally, Barro (2000) used the same short frequency data (he is focusing on ten-year intervals), but does not introduce a fixed effect. He finds that inequality is negatively associated with growth in the poorer countries, and positively in rich countries.

Banerjee and Duflo (2003b) investigate whether there is any reason to worry about the non-linearities that the theory suggests should be present. They find that when growth (or changes in growth) are regressed non-parametrically on changes in inequality, the relationship is an inverted U-shape. There is also a non-linear relationship between past inequality and the magnitudes of changes in inequality. Finally, there seems to be a negative relationship between growth rates and inequality lagged one period. These facts taken together, and in particular the non-linearities in these relationships (rather than the variation in samples or control variables), account for the different results obtained by different authors using different specifications.

[43] Forbes (2000) also corrects for the bias introduced by introducing a lagged variable in a fixed effect specification by using the GMM estimator developed by Arellano and Bond (1991).

6.3 Where do we go from here?

The discussion on functional form and identification, coupled with the empirical evidence of non-linearities even in very simple exercises, suggests that cross-country regressions are unlikely to be able to shed any meaningful light on the empirical relevance of models that integrate credit constraints and other imperfections of the credit markets. This is made worse by the poor quality of the aggregate data, despite the considerable efforts to produce consistent and reliable data sets. This contrasts with the increased availability of large, good quality, micro-economic data sets, which allow for testing specific hypotheses and derive credible identifying restrictions from theory and exogenous sources of variation. Throughout this chapter, we quoted many studies using micro-economic data which tested the micro-foundations for the models we discussed in this section.

Even a series of convincing micro-empirical studies will not be enough to give us an overall sense of how, together, they generate aggregate growth, the dynamics of income distribution, and the complex relationships between the two. The lessons of development economics will be lost to growth if they are not brought together in an aggregate context. In other words, it is not enough to use them to loosely motivate cross-sectional growth regression exercises—the discussion in this section is but an example of the misleading conclusions to which this can lead.

An alternative that seems likely to be much more fruitful is to try to build macroeconomic models that incorporate the features we discussed, and to use the results from the microeconomic studies as parameters in calibration exercises. The exercise we performed in section 5 of this chapter is an illustration of the kind of work that we can hope to do. There are a number of recent papers that in some ways go further in this direction than we have gone. In particular, Quadrini (1999) and Cagetti and Nardi (2003), for the U.S., and Paulson and Townsend (2003), for Thailand, try to calibrate a model with credit constraints to understand the correlation between wealth and the probability of becoming an entrepreneur. The paper by Buera (2003) mentioned above, emphasizes the fact that the long run correlation between wealth and entrepreneurship is weaker than the short run correlation, because as noted by Skiba (1978), Deaton (1992), Aiyagari (1994) and Carroll (1997), those who are credit constrained now but want to invest in the future have a very strong incentive to save. This, Buera points out, reduces the ultimate efficiency cost of imperfect credit markets, though in spite of this, the person with the median ability level and the median starting wealth loses about 18% of lifetime welfare because of the credit constraints. Caselli and Gennaioli (2002) offer a slightly different calibration: Like Buera, they are worried about the fact that with credit constraints the biggest firms may not be run by the best entrepreneurs. This can be a source of very large productivity losses in the short run. However, since the best entrepreneurs will make the most money, in the long run their firms would necessarily become the largest, unless they died young. They show that even with this limiting factor, reasonable death rates would imply a 20% loss of

productivity when we compare an economy without credit constraints with one that has them.

The calibrations so far have not attempted to see if the path of wealth distribution that results from calibrating this type of model matches the data. Our exercise above, for example, tries to match the distribution of firm sizes at a point of time, but says nothing about the path, while Buera does not try to match the data. The one exception is the papers by Robert Townsend and his collaborators based on Thai data (Jeong and Townsend (2003); Townsend and Ueda (2003)).

These papers, as well as those mentioned in the previous paragraphs, start from the assumption that every firm has a single, usually strictly concave, production technology. The only fixed cost comes from the fact that the firm needs an entrepreneur. As we saw above, this model does not do very well in terms of explaining the cross-sectional variation in the firm sector or the overall productivity gap, as compared to a model with a small number of alternative technologies and varying fixed costs. More generally, we need both a better empirical understanding of where the most important sources of inefficiency lie and better integration of this understanding when we assess the predictions of growth theory.

And perhaps above all, we need better growth theory: Our exercise at the beginning of this section was intended to advertise the possibility of a growth theory that does not assume aggregation. While we attempted to link the results to some relatively general properties of the production function, our analysis relies heavily on the fact that the inefficiency we assumed was in the credit market and that this took the form of a credit limit that was linear in wealth. One can easily imagine other ways for the credit market to be imperfect and other results from such models. Moreover, while the class of production technologies covered by our model was broader than usual, it does not include the (multiple-fixed-cost) technology that the previous section advocates.

There are, of course, other types of non-aggregative models: There are some examples of non-aggregative growth models that build on the inefficiency that comes from poorly functioning insurance markets.[44] There are also interesting attempts to build growth models that emphasize the fact that some people are favored by the government while others are not, and especially the fact that this changes over time in some predictable way (see Roland Benabou's contribution to this volume). Some interesting recent work has been done on the dynamic interplay between growth and political institutions (see the chapter by Acemoglu and Robinson in this volume) as well as between growth and social institutions (see Oded Galor's contribution to this volume, as well as Cole, Mailath and Postelwaithe (1992, 1998, 2001)). However, even more than in the case of the literature on credit markets and growth, it is not clear how much the insights from these models rely on specific details of how the environment or the imperfection

[44]See Banerjee and Newman (1991) for a theoretical model of non-aggregative growth based on imperfect insurance markets. Deaton and Paxson (1994) investigate some of empirical implications of this type of model using Taiwanese data. Krussel and Smith (1998) and Angeletos and Calvet (2003) are attempts to calibrate the impact of imperfect insurance on welfare and growth.

was modeled and to what extent they can be seen as robust properties of this entire class of models.

There are also areas where growth theory has not really reached: We have no models that, for example, incorporate reputation-building or learning into growth theory. The same can be said about the entire class of behavioral models of underinvestment.

Finally, there is the open question of whether we gain anything by building grand models that incorporate all these different reasons for inefficiency in a single model. To answer this we would need to assess whether the fact that different forms of inefficiency interact with each other has empirically important consequences.

This is an exciting time to think about growth. We are beginning to see the contours of a new vision, both more rooted in evidence and more ambitious in its theorizing.

References

Acemoglu, Daron, and Fabrizio Zilibotti (1994) 'Was Prometheus unbound by chance?' Mimeo, MIT

Acemoglu, Daron, and Joshua Angrist (2001) 'How large are human-capital externalities? Evidence from compulsory schooling laws.' In *NBER Macroeconomics Annual 2000, Vol. 15,* ed. Ben Bernanke and Kenneth Rogoff (Cambridge and London: MIT Press) pp. 9–59

Acemoglu, Daron, Simon Johnson, and James Robinson (2001) 'The colonial origins of comparative development: An empirical investigation.' *American Economic Review* 91(5), 1369–1401

Aghion, Philippe, and Patrick Bolton (1997) 'A trickle-down theory of growth and development with debt overhang.' *Review of Economic Studies* 64(2), 151–72

Aghion, Philippe, and Peter Howitt (1992) 'A model of growth through creative destruction.' *Econometrica* 60(2), 323–351

Aghion, Philippe, Peter Howitt, and David Mayer-Foulkes (2003) 'The effect of financial development on convergence: Theory and evidence.' Mimeo, Harvard University

Aiyagari, S. Rao (1994) 'Uninsured idiosyncratic risk and aggregate saving.' *Quarterly Journal of Economics* 109, 569–684

Aleem, Irfan (1990) 'Imperfect information, screening and the costs of informal lending: A study of a rural credit market in Pakistan.' *World Bank Economic Review* 3, 329–349

Alesina, Alberto, and Dani Rodrik (1994) 'Distributive politics and economic growth.' *Quarterly Journal of Economics* 109(2), 465–490

Anderson, Siwan, and Jean-Marie Baland (2002) 'The economics of roscas and intrahousehold resource allocation.' *Quarterly Journal of Economics* 117(3), 963–995

Angeletos, George-Marios, and Laurent Calvet (2003) 'Idiosyncratic production risk, growth and the business cycle.' Mimeo, MIT

Aportela, Fernando (1998) 'The effects of financial access on savings by low-income people.' Mimeo, MIT

Appleton, S., A. Bigsten-P. Collier-S. Dercon M. Fafchamps B. Gauthier J.W. Gunning A. Isaksson A. Oduro R. Oostendorp C. Pattillo M. Soderbom F. Teal, and A. Zeufack (1998) 'Rates of return on physical and human capital in Africa's manufacturing sector.' Working Paper WPS98.12, University of Oxford Centre for the Study of African Economies

Appleton, S., J. Hoddinott, and J. Knight (1996) 'Primary education as an input to post-primary education: A neglected benefit.' *Oxford Bulletin of Economics and Statistics* 58, 209–217

Appleton, Simon, Arne Bigsten, and Damiano Manda Kulundu (1999) 'Educational expansion and economic decline: Returns to education in Kenya 1978-1995.' Working Paper WPS99.06, University of Oxford Centre for the Study of African Economies

Arellano, Manuel, and Stephen Bond (1991) 'Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations.' *Review of Economic Studies* 58(2), 277–297

Banerjee, Abhijit V. (1992) 'A simple model of herd behavior.' *Quarterly Journal of Economics* 117(3), 797–817

——— (2001) 'The two poverties.' *Nordic Journal of Political Economy* 26(2), 129–141

——— (2003a) 'Contracting constraints, credit markets and economic development.' In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress, Vol. III,* ed. L. Hansen Dewatripont, M. and S. Turnovsky (Cambridge University Press)

——— (2003b) 'Educational policy and the economics of the family.' Mimeo, MIT

Banerjee, Abhijit V., and Andrew Newman (1991) 'Risk bearing and the theory of income distribution.' *Review of Economic Studies* 58(2), 211–235

——— (1993) 'Occupational choice and the process of development.' *Journal of Political Economy* 101(2), 274–298

Banerjee, Abhijit V., and Esther Duflo (2000) 'Reputation effects and the limits of contracting: A study of the indian software industry.' *Quarterly Journal of Economics* 115(3), 989–1017

——— (2003a) 'Do firms want to borrow more? Testing credit constraints using a directed lending program.' Working Paper 2003-5, Bureau for Research in Economic Analysis of Development

——— (2003b) 'Inequality and growth: What can the data say?' *Journal of Economic Growth* 8, 267–299

Banerjee, Abhijit V., and Kaivan Munshi (2004) 'How efficiently is capital allocated? Evidence from the knitted garment industry in Tirupur.' *Review of Economic Studies* 71(1), 19–42

Banerjee, Abhijit V., Esther Duflo, and Kaivan Munshi (2003) 'The (mis)allocation of capital.' *Journal of the European Economic Association* 1(2–3), 484–494

Banerjee, Abhijit V., Paul Gertler, and Maitreesh Ghatak (2002) 'Empowerment and efficiency: Tenancy reform in West Bengal.' *Journal of Political Economy* 110(2), 239–280

Bardhan, Pranab, and Christopher Udry (1999) *Development microeconomics* (Oxford, New York: Oxford University Press)

Barro, Robert J. (1974) 'Are government bonds net wealth?' *Journal of Political Economy* 82(6), 1095–1117

——— (1991) 'Economic growth in a cross section of countries.' *Quarterly Journal of Economics* 106(2), 407–443

——— (2000) 'Inequality and growth in a panel of countries.' *Journal of Economic Growth* 5(1), 5–32

Barro, Robert J., and Jong-Wha Lee (2000) 'International data on educational attainment updates and implications.' Working Paper 7911, National Bureau of Economic Research, September

Barro, Robert J., and Xavier Sala-I-Martin (1995) *Economic Growth* (New York: McGraw Hill)

Basta, S., Soekirman-D. Karyadi, and N. Scrimshaw (1979) 'Iron deficiency anemia and the productivity of adult males in Indonesia.' *American Journal of Clinical Nutrition* 32(4), 916–925

Becker, Gary (1981) *A Treatise on the Family* (Cambridge, MA: Harvard University Press)

Benabou, Roland (1996) 'Inequality and growth.' In *NBER Macroeconomics Annual 1996,* ed. Ben Bernanke and Julio J. Rotemberg (Cambridge and London: MIT Press) pp. 11–73

Benhabib, Jess, and A. Rustichini (1998) 'Social conflict and growth.' *Journal of Economic Growth* 1(1), 143–158

Benhabib, Jess, and Mark M. Spiegel (1994) 'The role of human capital in economic development: Evidence from aggregate cross-country data.' *Journal of Monetary Economics* 34(2), 143–174

Bennell, Paul (1996) 'Rates of return to education: Does the conventional pattern prevail in sub-saharan Africa?' *World Development* 24(1), 183–199

Besley, Timothy (1995) 'Savings, credit and insurance.' In *Handbook of Development Economics,* ed. J. Behrman and T.N. Srinivasan, vol. 3A (Amsterdam: Elsevier Science) chapter 6, pp. 2123–2207

Besley, Timothy, and Anne Case (1994) 'Diffusion as a learning process:Evidence from HYV cotton.' Discussion Paper 174, RPDS, Princeton University

Bigsten, Arne, Anders Isaksson Mans Soderbom-Paul Collier Et Al (2000) 'Rates of return on physical and human capital in Africa's manufacturing sector.' *Economic Development and Cultural Change* 48(4), 801–827

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch (1992) 'A theory of fads, fashion, custom, and cultural change as informational cascades.' *Journal of Political Economy* 100(5), 992–1026

Bils, Mark, and Peter Klenow (2000) 'Does schooling cause growth?' *American Economic Review* 90(5), 1160–1183

Binswanger, Hans, and Mark Rosenzweig (1986) 'Behavioural and material determinants of production relations in agriculture.' *Journal of Development Studies* 22(3), 503–539

Blanchflower, David, and Andrew Oswald (1998) 'What makes an entrepreneur?' *Journal of Labor Economics* 16(1), 26–60

Blom, Andreas, Dorte Verner, and Lauritz Holm-Nielsen (2001) 'Education, earnings, and inequality in Brazil, 1982-98: Implications for education policy.' Working Paper 2686, World Bank

Bottomley, A. (1963) 'The cost of administering private loans in underdeveloped rural areas.' *Oxford Economic Papers* 15(2), 154–163

Bourguignon, François, and Pierre-Andre Chiappori (1992) 'Collective models of household behavior.' *European Economic Review* 36, 355–364

Buera, Fancisco (2003) 'A dynamic model of entrepreneurship with borrowing constraints.' Mimeo, University of Chicago

Cagetti, Marco, and M. De Nardi (2003) 'Entrepreneurship, frictions and wealth.' Staff Report 324, Federal Reserve Bank of Minneapolis

Carroll, Christopher (1997) 'Buffer-stock saving and the life cycle/permanent income hypothesis.' *Quarterly Journal of Economics* 112, 1–56

Carvalho, Irineu (2000) 'Household income as a determinant of child labor and school enrollment in Brazil: Evidence from a social security reform.' Mimeo, MIT

Case, Anne, and Christina Paxson (2001) 'Mothers and others: Who invests in children's health?' *Journal of Health Economics* 20(3), 301–328

Case, Anne, Christina Paxson, and Joseph Ableidinger (2002) 'Orphans in Africa.' Working Paper 9213, National Bureau of Economic Research

Case, Anne, I-Fen Lin, and Sara McLanahan (2000) 'How hungry is the selfish gene?' *Economic Journal* 110(466), 781–804

Caselli, Francesco, and N. Gennaioli (2002) 'Dynastic management.' Mimeo, Harvard University

Coate, Stephen, and Martin Ravallion (1993) 'Reciprococity without commitment: Characterization and performance of informal insurance arrangements.' *Journal of Development Economics* 40, 1–24

Cole, Harold, George Mailath, and Andrew Postlewaite (1992) 'Social norms, savings behavior and growth.' *Journal of Political Economy* 100(6), 1092–1126

—— (1998) 'Class systems and the enforcement of social norms.' *Journal of Public Economics* 70(1), 5–35

—— (2001) 'Investment and concern for relative position.' *Review of Economic Design* 6, 241–261

Conley, Tim, and Christopher Udry (2003) 'Learning about a new technology: Pineapple in Ghana.' Mimeo, Northwestern University

Dasgupta, A. (1989) *Reports on Informal Credit Markets in India: Summary* (New Delhi: National Institute of Public Finance and Policy)

Dasgupta, Partha, and Debraj Ray (1986) 'Inequality as a determinant of malnutrition and unemployment: Theory.' *The Economic Journal* 96(384), 1011–1034

Deaton, Angus (1992) *Understanding Consumption* (Oxford: Oxford University Press)

—— (1997) *The Analysis of Household Surveys* (World Bank, International Bank for Reconstruction and Development)

Deaton, Angus, and Christina Paxson (1994) 'Intertemporal choice and inequality.' *Journal of Political Economy* 1-2(3), 437–467

Deininger, Klaus, and Lyn Squire (1996) 'A new data set measuring income inequality.' *World Bank Economic Review* 10, 565–591

Djankov, Simeon, Rafael La Porta Florencio Lopez-de-Silanes, and Andrei Shleifer (2002) 'The regulation of entry.' *Quarterly Journal of Economics* 117(1), 1–37

—— (2003) 'Courts.' *Quarterly Journal of Economics* 118(2), 453–518

Do, Toan, and Laksmi Iyer (2003) 'Land rights and economic development: Evidence from Vietnam.' Mimeo, Harvard Business School

Dreze, Jean, and Geeta Gandhi Kingdon (1999) 'School participation in rural India.' Working Paper 18, LSE Development Economics Discussion Paper Series

Duflo, Esther (2001) 'The medium run effects of educational expansion: Evidence from a large school construction program in Indonesia.' Working Paper 2003-2, Bureau for Research in Economic Analysis of Development. forthcoming, Journal of Developemnt Economics

—— (2003) 'Poor but rational.' Mimeo, MIT

Duflo, Esther, Michael Kremer, and James Robinson (2003) 'Understanding technology adoption: Fertilizer in Western Kenya, preliminary results from field experiments.' Mimeo, MIT

Duncan Thomas, Elizabeth Frankenberg, Jed Friedman Jean-Pierre Habicht, and Et Al (2003) 'Iron deficiency and the well being of older adults: Early results from a randomized nutrition intervention.' Mimeo, UCLA

Durlauf, Steven (1993) 'Nonergodic economic growth.' *Review of Economic Studies* 60(2), 349–366

Eaton, Jonathan, and Samuel Kortum (2001) 'Trade in capital goods.' *European Economic Review* 45(7), 1195–1235

Edmonds, Eric (2004) 'Does illiquidity alter child labor and schooling decisions? evidence from household responses to anticipated cash transfers in South Africa.' Working Paper 10265, National Bureau of Economic Research

Ellison, Glenn, and Drew Fudenberg (1993) 'Rules of thumbs for social learning.' *Journal of Political Economy* 101(4), 93–126

Ellison, Glenn, and Edward Glaeser (1997) 'Geographic concentration in U.S. manufacturing industries: A dartboard approach.' *Journal of Political Economy* 105(5), 889–927

—— (1999) 'Geographic concentration of industry: Does natural advantage explain agglomeration?' *American Economic Review* 89(2), 311–316

Eswaran, Mukesh, and Ashok Kotwal (1985) 'A theory of contractual structure in agriculture.' *American Economic Review* 75(3), 352–367

Fafchamps, Marcel (2000) 'Ethnicity and credit in African manufacturing.' *Journal of Development Economics* 61, 205–235

Fafchamps, Marcel, and Susan Lund (2003) 'Risk-sharing networks in rural Philippines.' *Review of Economic Studies* 71(2), 261–287

Fazzari, S., G. Hubbard, and B. Petersen (1988) 'Financing constraints and corporate investment.' *Brookings Papers on Economic Activity* 0(1), 141–195

Forbes, Kristin J. (2000) 'A reassessment of the relationship between inequality and growth.' *American Economic Review* 90(4), 869–887

Foster, Andrew D., and Mark R. Rosenzweig (1995) 'Learning by doing and learning from others: Human capital and technical change in agriculture.' *Journal of Political Economy* 103(6), 1176–1209

——— (1996) 'Technical change and human-capital returns and investments: Evidence from the green revolution.' *American Economic Review* 86(4), 931–953

——— (1999) 'Missing women, the marriage market and economic growth.' Mimeo, University of Pennsylvania

Foster, Andrew D., and Mark Rosenzweig (2000) 'Technological change and the distribution of schooling: Evidence from green revolution in India.' Mimeo, Brown University

Frazer, Garth (2001) 'Linking firms and workers: Heterogeneous labor and returns to education.' Mimeo, Yale University

Freeman, Richard (1986) 'Demand for education.' In *Handbook of Labor Economics,* ed. O. Ashenfelter and A. Layard, vol. 1 (Netherlands: Elsevier Publishers) chapter 6

Freeman, Richard, and Remco Oostendorp (2001) 'The occupational wages around the world data file.' *International Labour Review* 140(4), 379–401

Funkhouser, Edward (1998) 'Changes in the returns to education in Costa Rica.' *Journal of Development Economics* 57, 289–317

Galor, Oded, and Joseph Zeira (1993) 'Income distribution and macroeconomics.' *Review of Economic Studies* 60(1), 35–52

Gelos, R. Gaston, and Alejandro Werner (2002) 'Financial liberalization, credit constraints, and collateral: Investment in the Mexican manufacturing sector.' *Journal of Development Economics* 67(1), 1–27

Gertler, Paul, and Jonathan Gruber (2002) 'Insuring consumption against illness.' *American Economic Review* 92(1), 51–76

Gertler, Paul J., and Simone Boyce (2002) 'An experiment in incentive-based welfare: The impact of PROGESA on health in Mexico.' Mimeo, University of California, Berkeley

Ghatak, Maitreesh, Massimo Morelli, and Tomas Sjostrom (2001) 'Occupational choice and dynamic incentives.' *Review of Economic Studies* 68(4), 781–810

—— (2002) 'Credit rationing, wealth inequality, and allocation of talent.' Mimeo, London School of Economics

Ghate, Prabhu (1992) *Informal Finance: Some Findings from Asia* (Oxford; New York; Toronto and Hong Kong: Oxford University Press for the Asian Development Bank)

Goldstein, Markus, and Christopher Udry (1999) 'Agricultural innovation and resource management in ghana.' Mimeo, Yale University; Final Report to IFPRI under MP17

—— (2002) 'Gender, land rights and agriculture in Ghana.' Mimeo, Yale University

Greenwood, Jeremy, and Boyan Jovanovic (1999) 'The information technology revolution and the stock market.' *American Economic Review* 89(2), 116–22

Grossman, Gene, and Elhanan Helpman (1991) *Innovation and Growth in the Global Economy* (Cambridge, MA: MIT Press)

Gugerty, Mary Kay (2000) 'You can't save alone: Testing theories of rotating savings and credit associations.' Mimeo, Harvard University

Hart, O., and J. Moore (1994) 'A theory of debt based on inalienability of human capital.' *Quarterly Journal of Economics* 109, 841–879

Heston, Alan, Robert Summers, and Bettina Aten (2002) 'Penn World Table version 6.1.' Technical Report, Center for International Comparisons at the University of Pennsylvania

Hsieh, Chang-Tai (1999) 'Productivity growth and factor prices in east asia.' *American Economic Review* 89(2), 133–138

Hsieh, Chang-Tai, and Peter J. Klenow (2003) 'Relative prices and relative prosperity.' Mimeo, University of California, Berkeley

Jacoby, Hanan (2002) 'Is there an intrahousehold flypaper effect? Evidence from a school feeding program.' *Economic Journal* 112(476), 196–221

Jeong, Hyeok, and Robert Townsend (2003) 'Growth and inequality: Model evaluation based on an estimation-calibration strategy.' Mimeo, University of Southern California

Johnson, Simon, John McMillan, and Christopher Woodruff (2002) 'Property rights and finance.' *American Economic Review* 92(5), 1335–1356

Jovanovic, Boyan, and Peter Rousseau (2003) 'Specific capital and the division of rents.' Mimeo, New York University

Jovanovic, Boyan, and Rafael Rob (1997) 'Solow vs. Solow: Machine prices and development.' Working Paper 5871, National Bureau of Economic Research

Kanbur, Ravi (1979) 'Of risk taking and the personal distribution of income.' *Journal of Political Economy* 87, 769–797

Karlan, Dean (2003) 'Using experimental economics to measure social capital and predict financial decisions.' Mimeo, Princeton University

Kihlstrom, R., and J. Laffont (1979) 'A general equilibrium entrepreneurial theory of firm formation based on risk aversion.' *Journal of Political Economy* 87, 719–748

Knack, Stephen, and Philip Keefer (1995) 'Institutions and economic performance: Cross-country tests using alternative institutional measures.' *Economics and Politics* 7(3), 207–227

Kremer, Michael, and Edward Miguel (2003) 'The illusion of sustainability.' Mimeo, University of California, Berkeley

Kremer, Michael, Sylvie Moulin, and Robert Namunyu (2003) 'Decentralization: A cautionary tale.' Mimeo, Harvard University

Krishnan, Pramila, Tesfaye Gebre Selassie, and Stefan Dercon (1998) 'The urban labour market during structural adjustment: Ethiopia 1990-1997.' Working Paper WPS98.9, University of Oxford Centre for the Study of African Economies

Krueger, Alan, and Mikael Lindahl (1999) 'Education for growth: Why and for whom?' Mimeo, Princeton University

Krueger, Anne (1967) 'Factor endowments and per capital income differences among countries.' *Economic Journal* 78, 641–659

Krussel, Per, and Anthony Smith (1998) 'Income and wealth heterogeneity in the macroeconomy.' *Journal of Political Economy* 106(5), 867–896

La Porta, Rafael, Florencio Lopez-de-Silanes-Andrei Shleifer, and Robert Vishny (1998) 'Law and finance.' *Journal of Political Economy* 106(6), 1113–1155

Laffont, Jean-Jacques, and Mohamed Salah Matoussi (1995) 'Moral hazard, financial constraints and sharecropping in El Oulja.' *Review of Economic Studies* 62(3), 381–399

Laibson, David (1991) 'Golden eggs and hyperbolic discounting.' *Quarterly Journal of Economics* 62, 443–477

Levine, Ross, and David Renelt (1992) 'A sensitivity analysis of cross-country growth regressions.' *American Economic Review* 82(4), 942–963

Li, Hongyi, and Heng-fu Zou (1998) 'Income inequality is not harmful for growth: Theory and evidence.' *Review of Development Economics* 2(3), 318–334

Li, R., X. Chen H. Yan-P. Deurenberg L. Garby, and J.G. Hautvast (1994) 'Functional consequences of iron supplementation in iron-deficient female cotton workers in Beijing China.' *American Journal of Clinical Nutrition* 59, 908–913

Lloyd-Ellis, H., and D. Bernhardt (2000) 'Enterprise, inequality and economic development.' *Review of Economic Studies* 67(1), 147–169

Long, Sharon K. (1991) 'Do the school nutrition programs supplement household food expenditures?' *Journal of Human Resources* 26, 654–678

Lopez-Acevedo, Gladys (2001) 'Evolution of earnings and rates of returns to education in Mexico.' Working Paper 2691, World Bank Education Child Labor, Returns To Schooling Series

Loury, Glenn C. (1981) 'Intergenerational transfers and the distribution of earnings.' *Econometrica* 49(4), 843–867

Lucas, Robert (1990) 'Why doesn't capital flow from rich to poor countries?' *American Economic Review* 80(2), 92–96

Lundberg, Shelly, and Robert Pollak (1994) 'Noncooperative bargaining models of marriage.' *American Economic Review* 84(2), 132–137

—— (1996) 'Bargaining and distribution in marriage.' *Journal of Economic Perspectives* 10(4), 139–158

Mankiw, Gregory, David Romer, and David Weil (1992) 'A contribution to the empirics of economic growth.' *Quarterly Journal of Economics* 107(2), 407–437

Manski, Charles (1993) 'Identification of exogenous social effects: The reflection problem.' *Review of Economic Studies* 60, 531–542

Mason, A., and S. Khandker (1995) 'Household schooling decisions in tanzania.' Mimeo, World Bank

Matsuyama, Kiminori (2000) 'Endogenous inequality.' *Review of Economic Studies* 67(4), 743–759

____ (2003) 'On the rise and fall of class societies.' Mimeo, Northwestern University

Mauro, Paolo (1995) 'Corruption and growth.' *Quarterly Journal of Economics* 110(3), 681–712

McKenzie, David, and Christopher Woodruff (2003) 'Do entry costs provide an empirical basis for poverty traps? Evidence from Mexican microenterprises.' Working Paper 2003-20, Bureau for Research in Economic Analysis of Development

McKinsey Global Institute (2001) 'India: The growth imperative.' Report, McKinsey Global Institute

Miguel, Edward, and Michael Kremer (2003) 'Networks, social learning, and technology adoption: The case of deworming drugs in Kenya.' Mimeo, University of California, Berkeley

____ (2004) 'Worms: Identifying impacts on education and health in the presence of treatment externalities.' *Econometrica* 72(1), 159–218

Mookherjee, Dilip, and Debraj Ray (2002) 'Contractual structure and wealth accumulation.' *American Economic Review* 92(4), 818–849

____ (2003) 'Persistent inequality.' *Review of Economic Studies* 70(2), 369–393

Morduch, Jonathan (1993) 'Risk production and saving: Theory and evidence from Indian households.' Mimeo, Harvard University

Munshi, Kaivan (2003) 'Social learning in a heterogeneous population: Technology diffusion in the Indian green revolution.' Mimeo, University of Pennsylvania

Munshi, Kaivan, and Jacques Myaux (2002) 'Development as a process of social change: An application to the fertility transition.' Mimeo, Brown University

Munshi, Kaivan, and Mark Rosenzweig (2004) 'Traditional institutions meet the modern world: Caste, gender, and schooling choice in a globalizing economy.' Mimeo, Brown University

Murphy, Kevin, Andrei Shleifer, and Robert Vishny (1989) 'Industrialization and the big push.' *Journal of Political Economy* 97(5), 1003–1026

—— (1995) 'The allocation of talent: Implications for growth.' In *The economic analysis of rent seeking,* ed. Robert Tollison and Roger Congleton, vol. 49 of *Elgar Reference Collection: International Library of Critical Writings in Economics* (Ashgate; Elgar) pp. 301–328

Murshid, K. (1992) 'Informal credit markets in Bangladesh agriculture: Bane or boon?' In 'Sustainable Agricultural Development: The role of international cooperation: Proceedings of the 21st International Conference of Agricultural Economists, held at Tokyo, Japan, 22-29 August 1991' (Aldershot, U.K.: Dartmouth; Brookfield, Vt: Ashgate) pp. 657–68

Mwabu, G., and T. Schultz (1995) 'Wage premia for education and location by gender and race in South Africa.' Mimeo, Yale University

Newman, Andrew (1995) 'Risk-bearing and 'Knightian' entrepreneurship.' Mimeo, Columbia University

Olley, G. Steven, and Ariel Pakes (1996) 'The dynamics of productivity in the telecommunications equipment industry.' *Econometrica* 64(6), 1263–1297

on Macroeconomics, Commission, and Health (2001) *Macroeconomics and health: investing in health for economic development: Report* (Geneva: World Health Organization)

Parente, Stephen, and Edward Prescott (1994) 'Barriers to technology adoption and development.' *Journal of Political Economy* 102(2), 298–321

—— (2000) *Barriers to Riches: Walras-Pareto Lectures, vol. 3* (Cambridge, MA: MIT Press)

Paulson, Anna, and Robert Townsend (2003) 'Entrepreneurship and financial constraints in thailand.' Mimeo, Northwestern University, forthcoming in Journal of Corporate Finance

Persson, Torsten, and Guido Tabellini (1991) 'Is inequality harmful for growth? Theory and evidence.' *American Economic Review* 48, 600–621

Piketty, Thomas (1997) 'The dynamics of the wealth distribution and the interest rate with credit rationing.' *The Review of Economic Studies* 64(2), 173–189

Powell, Christine, Sally Grantham-McGregor, and M. Elston (1983) 'An evaluation of giving the Jamaican government school meal to a class of children.' *Human Nutrition: Clinical Nutrition* 37C, 381–388

Powell, Christine, Sally Walker-Susan Chang, and Sally Grantham-McGregor (1998) 'Nutrition and education: A randomized trial of the effects of breakfast in rural primary school children.' *American Journal of Clinical Nutrition* 68, 873–879

Psacharopoulos, George (1973) *Returns to Education: An International Comparison* (San Francisco: Jossy Bass-Elsevier)

—— (1985) 'Returns to education: A further international update and implications.' *Journal of Human Resources* 20(4), 583–604

—— (1994) 'Returns to investments in education: a global update.' *World Development* 22(9), 1325–1343

—— (2002) 'Returns to investment in education: A global update.' *World Development* 22(9), 1325–1343

Qian, Nancy (2003) 'Missing women and the price of tea in china.' Mimeo, MIT

Quadrini, Vincenco (1999) 'The importance of entrepreneurship for wealth concentration and mobility.' *Review of Income and Wealth* 45, 1–19

Rauch, James (1993) 'Productivity gains from geographic concentration of human capital: Evidence from the cities.' *Journal of Urban Economics* 34(3), 380–400

Ravallion, Martin, and Shubham Chaudhuri (1997) 'Risk and insurance in village India: Comment.' *Econometrica* 65(1), 171–184

Ray, Debraj (1998) *Development Economics* (Princeton, N.J.: Princeton University Press)

—— (2003) 'Aspirations, poverty and economic change.' Mimeo, New York University

Restuccia, Diego, and Carlos Urrutia (2001) 'Relative prices and investment rates.' *Journal of Monetary Economics* 47(1), 93–121

Rosenstein-Rodan, Paul N. (1943) 'Problems of industrialization of eastern and south-eastern Europe.' *Economic Journal* 53, 202–211

Rosenzweig, Mark R., and Hans Binswanger (1993) 'Wealth, weather risk and the composition and profitability of agricultural investments.' *Economic Journal* 103(416), 56–78

Rosenzweig, Mark R., and Kenneth I. Wolpin (1993) 'Credit market constraints, consumption smoothing, and the accumulation of durable production assets in low-income countries: Investments in bullocks in India.' *Journal of Political Economy* 101(21), 223–244

Rosenzweig, Mark R., and T. Paul Schultz (1982) 'Market opportunities, genetic endowments, and intrafamily resource distribution: Child survival in rural India.' *American Economic Review* 72(4), 803–815

Sala-I-Martin, Xavier (1997) 'I just ran four million regressions.' Working Paper 6252, National Bureau of Economic Research

Schultz, T. Paul (1994) 'Human capital investment in women and men.' Occasioinal Paper 44, International Center for Economic Growth

—— (2001) 'School subsidies for the poor: Eevaluating the Mexican PROGRESA poverty program.' Mimeo, Yale University, forthcoming in Journal of Development Economics

Shaban, Radwan (1987) 'Testing between competing models of sharecropping.' *Journal of Political Economy* 95(5), 893–920

Skiba, A. K. (1978) 'Optimal growth with a convex-concave production function.' *Econometrica* 46, 527–539

Stiglitz, J. (1969) 'The effects of income, wealth, and capital gains taxation on risk-taking.' *Quarterly Journal of Economics* 83(2), 263–283

—— (1974) 'Incentives and risk sharing in sharecropping.' *Review of Economic Studies* 41(2), 219–255

Strauss, John, and Duncan Thomas (1995) 'Human resources: Empirical modeling of household and family decisions.' In *Handbook of Development Economics,* ed. Jere Behrman and T.N. Srinivasan, vol. 3A (Amsterdam: North Holland) chapter 34, pp. 1885–2023

Strauss, John, and Duncan Thomas (1998) 'Health, nutrition, and economic development.' *Journal of Economic Literature* 36(2), 766–817

Svensson, Jakob (1998) 'Investment, property rights and political instability: Theory and evidence.' *European Economic Review* 42(7), 1317–1341

Thomas, Duncan (2001) 'Health, nutrition and economic prosperity: A microeconomic perspective.' Working Paper, Bulletin of the World Working Paper 7, Working Group 1, WHO Commission on Macroeconomics and Health

Thomas, Duncan, and Elizabeth Frankenberg (2002) 'Health, nutrition and prosperity: A microeconomic perspective.' *Bulletin of the World Health Organization* 80(2), 106–113

Timberg, Thomas, and C. V. Aiyar (1984) 'Informal credit markets in India.' *Economic Development and Cultural Change* 33(1), 43–59

Tirole, Jean (1996) 'A theory of collective reputations (with applications to the persistance of corruption and to firm quality).' *Review of Economic Studies* 63(1), 1–22

Townsend, Robert (1994) 'Risk and insurance in village India.' *Econometrica* 62(4), 539–591

—— (1995) 'Financial systems in Northern Thai villages.' *Quarterly Journal of Economics* 110(4), 1011–1046

Townsend, Robert, and Kenichi Ueda (2003) 'Financial deepening, inequality, and growth: A model-based quantitative evaluation.' Mimeo, University of Chicago

Udry, Christopher (1990) 'Credit markets in Northern Nigeria: Credit as insurance in a rural economy.' *World Bank Economic Review* 4(3), 251–69

—— (1996) 'Gender, agricultural production, and the theory of the household.' *Journal of Political Economy* 101(5), 1010–1045

—— (2003) 'A note on the returns to capital in a developing country.' Mimeo, Yale University

United Nations Development Program (2001) *Human development report 2001: Making new technologies work for human development* (Oxford and New York: Oxford University Press)

Vere, James (2001) 'Education, technology and the wage structure in Taiwan, 1979-1998.' Mimeo, Princeton University

Vermeersch, Christel (2002) 'School meals, educational achievement and school competition: Evidence from a randomized evaluation.' Mimeo, Harvard University

Young, Alwyn (1995) 'The tyranny of numbers: Confronting the statistical realities of the East Asian growth experience.' *Quarterly Journal of Economics* 110(3), 641–680

Zeldes, Stephen (1989) 'Consumption and liquidity constraints: An empirical investigation.' *Journal of Political Economy* 97(2), 305–346

Table 1: Returns to Education

Variable	Mincerian returns				log(teacher salary)	direct costs/benefits	total costs/benefits
Sample	Psacharopoulous	Psacharopoulous, extended	Psacharopoulos, high quality	Psacharopoulos, high quality			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	16.40	13.01	11.04	9.65	2.24	4.09	21.43
	(2.6)	(1.35)	(1.14)	(.46)	(.15)	(.21)	(1.63)
Mean years of schooling	-0.72	-0.47	-0.27				
	(.3)	(.16)	(.14)				
GDP/capita (*1000)				-0.084		-0.034	-0.155
				(.039)		(.019)	(.147)
lgdp					0.79		
					(.02)		
n	37	70	62	62	532	61	61
r^2	0.139	0.106	0.062	0.072	0.7902	0.05	0.018

Source: The data on returns to education was compiled starting from Psacharopoulous (2000) and extedended by surveying the literature. Appendix table 1 lists the data and the sources. The data on teacher salary is from Freeman and Oosterkberke. The data on pupil teacher ratio is from UNESCO (2000).

Appendix table 1: rate of returns to education and years of schooling.

Country	Continent	year	mincerian returns	years of schooling (Psacharopoulos)	Years of schooling (world bank)	Source	Data rating (bennel)	Additions to Psacharopoulos data
Argentina	South America	1989	10.3	9.1	8.83	Psacharopoulos (1994)		
Australia	Australia	1989	8	.	10.92	Cohn and Addison (1998)		
Austria	Europe	1993	7.2	.	8.35	Fersterer and Winter-Ebmer (1999)		
Bolivia	South America	1993	10.7	.	5.58	Patrinos (1995)		
Botswana	Africa	1979	19.1	3.3	6.28	Psacharopoulos (1994)	Poor	
Brazil	South America	1998	12.21	5.3	4.88	Verner (2001)		Added
Burkina Faso	Africa	1980	9.6	.	.	Psacharopoulos (1994)	Poor	
Cameroon*	Africa	1995	5.96		3.54	Appleton et al (1999)		Added
Canada	North America	1989	8.9	.	11.62	Cohn (1997)		
Chile	South America	1989	12	8.5	7.55	Psacharopoulos (1994)		
China	Asia	1993	12.2	.	6.36	Hossain (1997)		
Colombia	South America	1989	14	8.2	5.27	Psacharopoulos (1994)		
Costa Rica	South America	1992	8.50	.	6.05	Funkhouser (1998)		Added
Cote d'Ivoire	Africa	1987	13.10	6.9	.	Schultz (1994)	Poor	Added
Cyprus	Europe	1994	5.2	.	9.15	Menon (1995)		
Denmark	Europe	1990	4.5	.	9.66	Christensen and Westergard-Nielsen (1999)		
Dominican Re	South America	1989	9.4	8.8	4.93	Psacharopoulos (1994)		
Ecuador	South America	1987	11.8	9.6	6.41	Psacharopoulos (1994)		
Egypt	Africa	1997	7.80	.	5.51	Wahba (2000)		
El Salvador	South America	1992	7.6	.	5.15	Funkhouser (1996)		
Estonia	Europe	1994	5.4	10.9	.	Kroncke (1999)		
Ethiopia	Africa	1997	3.28	6	.	Krishnan, Selasie, Dercon (1989)	Poor	Added
Finland	Europe	1993	8.2	.	9.99	Asplund (1999)		
France	Europe	1977	10	6.2	7.86	Psacharopoulos (1994)		
Germany	Europe	1988	7.7	.	10.2	Cohn and Addison (1998)		
Ghana	Africa	1999	8.80	9.7	3.89	Frazer (1998)		Added
Greece	Europe	1993	7.6	.	8.67	Magoula and Psacharopoulos (1999)		
Guatemala	South America	1989	14.9	4.3	3.49	Psacharopoulos (1994)		
Honduras	South America	1991	9.3	.		Funkhouser (1996)		
Hong Kong	Asia	1981	6.1	9.1	4.8	Psacharopoulos (1994)		
Hungary	Europe	1987	4.3	11.3	9.13	Psacharopoulos (1994)		
India	Asia	1995	10.6	.	5.06	Kingdon (1998)		
Indonesia	Asia	1995	7	8	4.99	Duflo (2000)		
Iran	Asia	1975	11.6	.	5.31	Psacharopoulos (1994)	Poor	
Israel	Asia	1979	6.4	11.2	9.6	Psacharopoulos (1994)	Poor	
Italy	Europe	1987	2.7	.	7.18	Brunello, Comi and Lucifora (1999)		
Jamaica	South America	1989	28.8	7.2	5.26	Psacharopoulos (1994)	Poor	

Japan	Asia	1988	13.2	.	9.47	Cohn and Addison (1998)		
Kenya	Africa	1995	11.39	8	4.2	Appleton et al (1998)		Added
Korea	Asia	1986	13.5	8	10.84	Ryoo, Nam and Carnoy (1993)		
Kuwait	Asia	1983	4.5	8.9	7.05	Psacharopoulos (1994)	Poor	
Malaysia	Asia	1979	9.4	15.8	6.8	Psacharopoulos (1994)		
Mexico	South America	1997	35.31	.	7.23	Lopez-Acevedo (2001)		Added
Morocco	Africa	1970	15.8	2.9 .		Psacharopoulos (1994)	Poor	
Nepal	Asia	1999	9.7	3.9	2.43	Parajuli (1999)		
Netherlands	Europe	1994	6.4	.	9.36	Hartog, Odink and Smits (1999)		
Nicaragua	South America	1996	12.1	.	4.58	Belli and Ayadi (1998)		
Norway	Europe	1995	5.5	.	11.85	Barth and Roed (1999)		
Pakistan	Asia	1991	15.4	.	3.88	Katsis, Mattson and Psacharopoulos (1998)		
Panama	South America	1990	13.7	9.2	8.55	Psacharopoulos (1994)		
Paraguay	South America	1990	11.5	9.1	6.18	Psacharopoulos (1994)		
Peru	South America	1990	8.1	10.1	7.58	Psacharopoulos (1994)		
Philippines	South America	1998	12.6	8.8	8.21	Schady (2000)		
Poland	Europe	1996	7	.	9.84	Nesterova and Sabirianova (1998)		
Portugal	Europe	1991	8.6	.	5.87	Cohn and Addison (1998)		
Puerto Rico	South America	1989	15.1	. .		Griffin and Cox Edwards (1993)		
Russian Federa	Europe	1996	7.2	11.7 .		Nesterova and Sabirianova (1998)		
Singapore	Asia	1998	13.1	9.5	7.05	Sakellariou (2001)		
South Africa*	Africa	1993	10.27	7.1	6.14	Mwabu and Schultz (1995)		Added
Spain	Europe	1991	7.2	.	7.28	Mora (1999)		
Sri Lanka	Asia	1981	7	4.5	6.87	Psacharopoulos (1994)		
Sudan	Africa	1989	9.3	10.2	2.14	Cohen and House (1994)		
Sweden	Europe	1991	5	.	11.41	Cohn and Addison (1998)		
Switzerland	Europe	1991	7.5	.	10.48	Weber and Wolter (1999)		
Taiwan	Asia	1998	19.01	9		Vere (2001)		Added
Tanzania*	Africa	1991	13.84	.	2.71	Mason and Kandker (1995)	Poor	Added
Thailand	Asia	1989	11.5	.	6.5	Patrinos (1995)		
Tunisia	Africa	1980	8	4.8	5.02	Psacharopoulos (1994)	Poor	
Uganda	Africa	1992	5.94		3.51	Appleton et al (1996)		Added
United Kingdo	Europe	1987	6.8	11.8	9.42	Psacharopoulos (1994)		
United States	North America	1995	10	.	12.05	Rouse (1999)		
Uruguay	South America	1989	9.7	9	7.56	Psacharopoulos (1994)		
Venezuela	South America	1992	9.4	.	6.64	Psacharopoulos and Mattson (1998)		
Vietnam	Asia	1992	4.8	7.9 .		Moock, Patrinos and Venkataraman (1998)		
Yugoslavia	Europe	1986	4.8	. .		Bevc (1993)		

Zambia*	Africa	1995	10.65	5.46	Appleton et al (1999)	Added
Zimbabwe*	Africa	1994	5.57	5.35	Appleton et al (1999)	Added

Notes: This table updates Psacharopoulos (2002). The last column indicate which rate of returns were added by us
The data rating quality is from Bennell (1996), and concerns only African Countries

Growth Econometrics

Steven N. Durlauf, Paul A. Johnson and Jonathan R. W. Temple

Final Draft: October 22, 2004

Durlauf thanks the University of Wisconsin and John D. and Catherine T. MacArthur Foundation for financial support. Johnson thanks the Department of Economics, University of Wisconsin for its hospitality in Fall 2003, during which part of this chapter was written. Temple thanks the Leverhulme Trust for financial support under the Philip Leverhulme Prize Fellowship scheme. Ritesh Banerjee, Ethan Cohen-Cole, Giacomo Rondina and Lisa Wong have provided excellent research assistance. Finally, we thank Gordon Anderson, William Brock and Stephen Bond for useful discussions.

Growth Econometrics

Abstract

This paper provides a survey and synthesis of econometric tools that have been employed to study economic growth. While these tools range across a variety of statistical methods, they are united in the common goals of first, identifying interesting contemporaneous patterns in growth data and second, drawing inferences on long-run economic outcomes from cross-section and temporal variation in growth. We describe the main stylized facts that have motivated the development of growth econometrics, the major statistical tools that have been employed to provide structural explanations for these facts, and the primary statistical issues that arise in the study of growth data. An important aspect of the survey is attention to the limits that exist in drawing conclusions from growth data, limits that reflect model uncertainty and the general weakness of available data relative to the sorts of questions for which they are employed.

Steven N. Durlauf
Department of Economics
University of Wisconsin
1180 Observatory Drive
Madison, WI 53706-1393
United States

Paul A. Johnson
Department of Economics
Vassar College
124 Raymond Avenue
Poughkeepsie, NY 12064-0708
United States

Jonathan R. W. Temple
Department of Economics
University of Bristol
8 Woodland Road
Bristol BS8 1TN
United Kingdom

The totality of our so-called knowledge or beliefs, from the most causal matters of geography and history to the profoundest laws of atomic physics…is a man-made fabric which impinges on experience only along the edges…total science is like a field of force whose boundary conditions are experience…A conflict with experience on the periphery occasions readjustments in the interior of the field. Reevaluation of some statements entails reevaluation of others, because of their logical interconnections…But the total field is so underdetermined by its boundary conditions, experience, that there is much latitude of choice as to what statements to reevaluate in the light of any single contrary experience.

W. V. O. Quine[1]

I. Introduction

The empirical study of economic growth occupies a position that is notably uneasy. Understanding the wealth of nations is one of the oldest and most important research agendas in the entire discipline. At the same time, it is also one of the areas in which genuine progress seems hardest to achieve. The contributions of individual papers can often appear slender. Even when the study of growth is viewed in terms of a collective endeavor, the various papers cannot easily be distilled into a consensus that would meet standards of evidence routinely applied in other fields of economics.

A traditional defense of empirical growth research would be in terms of expected payoffs. Each time an empirical growth paper is written, the probability of gaining genuine understanding may be low, but the payoff to that understanding is potentially vast. But even this argument relies on being able to discriminate between the status of different pieces of evidence – the good, the bad and the ugly – and this process of discrimination carries many difficulties of its own.

Rodriguez and Rodrik (2001) begin their skeptical critique of evidence on trade policy and growth with an apt quote from Mark Twain: "It isn't what we don't know that kills us. It's what we know that ain't so." This point applies with especial force in the identification of empirically salient growth determinants. As illustrated in Appendix 2 of this chapter, approximately as many growth determinants have been proposed as there are countries for which data are available. It is hard to believe that all these determinants are

[1] "Two Dogmas of Empiricism," *Philosophical Review*, 1951.

central, yet the embarrassment of riches also makes it hard to identify the subset that truly matters.

There are other respects in which it is difficult to reconcile alternative empirical studies, including the functional form posited for the growth process. An important distinction between the neoclassical growth model of Solow (1956) and Swan (1956) and many of the models that have been produced in the endogenous growth theory literature launched by Romer (1986) and Lucas (1988) is that the latter can require the specification of a nonlinear data generating process. But researchers have not yet agreed on the empirical specification of growth nonlinearities, or the methods that should be used to distinguish neoclassical and endogenous growth models empirically.

These and other difficulties inherent in the empirical study of growth have prompted the field to evolve continuously, and to adopt a wide range of methods. We argue that a sufficiently rich set of statistical tools for the study of growth have been developed and applied that they collectively define an area of growth econometrics. This chapter is designed to provide an overview of the current state of this field. The chapter will both survey the body of econometric and statistical methods that have been brought to bear on growth questions and provide some assessments of the value of these tools.

Much of growth econometrics reflects the specialized questions that naturally arise in growth contexts. For example, statistical tools are often used to draw inferences about long-run outcomes from contemporary behaviors. This is most clearly seen in the context of debates over economic convergence; as discussed below, many of the differences between neoclassical and endogenous growth perspectives may be reduced to questions concerning the long-run effects of initial conditions. The available growth data typically span at most 140 years (and many fewer if one wants to work with a data set that nontrivially spans countries outside Western Europe and the United States) and the use of these data to examine hypotheses about long-run behavior can be a difficult undertaking. Such exercises lead to complicated questions concerning how one can identify the steady-state behavior of a stochastic process from observations along its transition path.

As we have already mentioned, another major and difficult set of growth questions involves the identification of empirically salient determinants of growth when

the range of potential factors is large relative to the number of observations. Model uncertainty is in fact a fundamental problem facing growth researchers. Individual researchers, seeking to communicate the extent of support for particular growth determinants, typically emphasize a single model (or small set of models) and then carry out inference as if that model had generated the data. Standard inference procedures based on a single model, and which are conditional on the truth of that model, can grossly overstate the precision of inferences about a given phenomenon. Such procedures ignore the uncertainty that surrounds the validity of the model. Given that there are usually other models that have strong claims on our attention, the standard errors can understate the true degree of uncertainty about the parameters, and the choice of which models to report can appear arbitrary. The need to properly account for model uncertainty naturally leads to Bayesian or pseudo-Bayesian approaches to data analysis.[2]

Yet another set of questions involves the characterization of interesting patterns in a data set comprised of objects as complex and heterogeneous as countries. Assumptions about parameter constancy across units of observation seem particularly unappealing for cross-country data. On the other hand, much of the interest in growth economics stems precisely from the objective of understanding the distribution of outcomes across countries. The search for data patterns has led to a far greater use of classification and pattern recognition methods, for example, than appears in other areas of economics. Here and elsewhere, growth econometrics has imported a range of methods from statistics, rather than simply relying on the tools of mainstream econometrics.

Whichever techniques are applied, the weakness of the available data represents a major constraint on the potential of empirical growth research. Perhaps the main obstacle to understanding growth is the small number of countries in the world. This is a problem for the obvious reason (a fundamental lack of variation or information) but also because it limits the extent to which researchers can address problems such as measurement error and parameter heterogeneity. Sometimes the problem is stark: imagine trying to infer the consequences of democracy for growth in poorer countries. Because the twentieth century provided relatively few examples of stable, multi-party democracies among the

[2]See Draper (1995) for a general discussion of model uncertainty and Brock, Durlauf, and West (2003) for discussion of its implications for growth econometrics.

poorer nations of the world, statistical evidence can make only a limited contribution to this debate, unless one is willing to make exchangeability assumptions about nations that would seem not to be credible.[3]

With a larger group of countries to work with, many of the difficulties that face growth researchers could be addressed in ways that are now standard in the microeconometrics literature. For example, the well known concerns expressed by Harberger (1987), Solow (1994) and many others about assuming a common linear model for a set of very different countries could, in principle, be addressed by estimating more general models that allow for heterogeneity. This can be done using interaction terms, nonlinearities or semiparametric methods, so that the marginal effect of a given explanatory variable can differ across countries or over time. The problem is that these solutions will require large samples if the conclusions are to be robust. Similarly, some methods for addressing other problems, such as measurement error, are only useful in samples larger than those available to growth researchers. This helps to explain the need for new statistical methods for growth contexts, and why growth econometrics has evolved in such a pragmatic and eclectic fashion.

One common response to the lack of cross-country variation has been to draw on variation in growth and other variables over time, primarily using panel data methods. Many empirical growth papers are now based on the estimation of dynamic panel data models with fixed effects. Our survey will discuss not only the relevant technical issues, but also some issues of interpretation that are raised by these studies, and especially their treatment of fixed effects as nuisance parameters. We also discuss the merits of alternatives. These include the before-and-after studies of specific events, such as stock market liberalizations or democratizations, which form an increasingly popular method for examining certain hypotheses. The correspondence between these studies and the microeconometric literature on treatment effects helps to clarify the strengths and limitations of the event-study approach, and of cross-country evidence more generally.

[3]See Temple (2000b) and Brock and Durlauf (2001a) for a conceptual discussion of this issue.

Despite the many difficulties that arise in empirical growth research, we believe some progress has been made. Researchers have uncovered stylized facts that growth theories should endeavor to explain, and developed methods to investigate the links between these stylized facts and substantive economic arguments. We would also argue that an important contribution of growth econometrics has been the clarification of the limits that exist in employing statistical methods to address growth questions. One implication of these limits is that narrative and historical approaches (Landes (1998) and Mokyr (1992) are standard and valuable examples) have a lasting role to play in empirical growth analysis. This is unsurprising given the importance that many authors ascribe to political, social and cultural factors in growth, factors that often do not readily lend themselves to statistical analysis.[4] For these reasons, Willard Quine's classic statement of the underdetermination of theories by data, cited at the beginning of this chapter, seems especially relevant to the study of growth.

The chapter is organized as follows. Section II describes a set of stylized facts concerning economic growth. These facts constitute the objects that formal statistical analysis has attempted to explain. Section III describes the relationship between theoretical growth models and econometric frameworks for growth, with a primary focus on cross-country growth regressions. Section IV discusses the convergence hypothesis. Section V describes methods for identifying growth determinants, and a range of questions concerning model specification and evaluation are addressed. Section VI discusses econometric issues that arise according to whether one is using cross-section, time series or panel data, and also examines the issue of endogeneity in some depth. Section VII evaluates the implications of different data and error properties for growth analysis. Section VIII concludes with some thoughts on the progress made thus far, and possible directions for future research.

[4]Narrative approaches can, of course, be subjected to criticisms every bit as severe as apply to quantitative studies. Similarly, efforts to study qualitative growth ideas using formal tools can go awry; see Durlauf (2002) for criticism of efforts to explain growth and development using the idea of social capital.

II. Stylized facts

In this section we describe some of the major features of cross-country growth data. Our goal is to identify some of the salient cross-section and intertemporal patterns that have motivated the development of growth econometrics. Section II.i makes some general observations on growth in the very long-run. Section II.ii discusses the main data set used to study growth since 1960. Section II.iii describes general facts about differences in output per worker across countries. Section II.iv extends this discussion by focusing on growth miracles and disasters. Basic facts concerning convergence are reported in Section II.v. In Section II.vi we describe the general slowdown in growth over the last two decades. Section II.vii extends this discussion by considering the question of predictability of growth rates over time. Section II.viii identifies growth differences across levels of development and across geographic regions. In Section II.ix, we characterize some aspects of stagnation and volatility. Section II.x draws some general conclusions about the basic growth facts.

i. a long-run view

Taking a long view of economic history, a central fact concerning aggregate economic activity across countries is the massive divergence in living standards that has occurred over the last several centuries. A snapshot of the world in 1700 would show all countries to be poor, if their living standards were assessed in today's terms. Over the course of the 18^{th} and 19^{th} centuries, growth rates increased slightly in the UK and other countries in Western Europe. Annual growth rates appear to have remained low, by modern standards, even in the midst of the Industrial Revolution; but because this growth was sustained over time, GDP per capita steadily rose. The outcome was that the UK, some other countries in Western Europe, and then the USA gradually advanced further ahead of the rest of the world.

What was happening elsewhere? As Pritchett (1997) argues, even in the absence of national accounts data, we can be almost certain that rapid productivity growth was never sustained in the poorer regions of the world. The argument proceeds by

extrapolating backwards from their current levels of GDP per capita, using a fast growth rate. This quickly implies earlier levels of income that would be too low to support human life.

ii. data after 1960

Today's overall inequality across countries is partly the legacy of rapid growth in a small group of Western economies, and its absence elsewhere. But there have been important deviations from this general pattern. Since the 1960s, some developing countries have grown at rates that are unprecedented, at least based on the experiences of the advanced economies of Europe and North America. The tiger economies of East Asia have seen GDP per worker grow at around 5% a year, or even faster, for the best part of forty years. A country that grows at such rates over forty years will see GDP per worker rise more than sevenfold, as in the case of Hong Kong, Singapore, South Korea and Taiwan.

In the rest of this section, we describe these patterns in more detail. As with most of the empirical growth literature, we will focus on the period after 1960, the point at which national accounts data start to become available for a larger group of countries.[5] Our calculations use version 6.1 of the Penn World Table (PWT) due to Heston, Summers, and Aten (2002). They have constructed measures of real GDP that adjust for international differences in price levels, and are therefore more comparable across space than measures based on market exchange rates.[6]

For the purposes of our analysis, the "world" will consist of 102 countries, those with data available in PWT 6.1 and with populations of at least 350,000 in the year 1960. These 102 countries account for a large share of the world's population. The most important missing countries are economies in Eastern Europe that were centrally planned for much of the period. Because of its enormous population, collectivist China is included in the sample, but is a country for which output measurement is especially difficult. In a

[5]Another reason for this starting point is that many colonies did not gain independence until the 1960s.

[6]For more discussion of the PWT data, and further references, see Temple (1999).

small number of cases, data for GDP per worker for 2000 are extrapolated from preceding years using growth rates for the early and mid-1990s. The Appendix gives more details of the sample, and the extrapolation procedure.

Throughout, we use data on GDP per worker. Most formal growth models are based on production functions, and their implications relate more closely to GDP per worker than GDP per capita. Jones (1997) provides another justification for this choice. When there is an unmeasured non-market sector, such as subsistence agriculture, GDP per worker could be a more accurate index of average productivity than GDP per capita.

The paths of GDP per worker and GDP per capita will diverge when there are changes in the ratio of workers to population, which is one form of participation rate. There has been an upwards trend in these participation rates where such rates were originally low, while at the upper end of the distribution participation has been stable.[7] For a sample of 90 countries with available data, the median participation rate rose from 41% to 45% between 1960 and 2000. There was a sharp increase at the 25th percentile (from 33% to 40%) but very little change at the 75th percentile. This pattern suggests that growth in GDP per capita has usually been close to growth in GDP per worker, except for the countries that started with low participation rates.

There is an important point to bear in mind, when interpreting our later tables and graphs, and those found elsewhere in the literature. Our unit of observation is the country. In one sense this is clearly an arbitrary way to divide the world's population, but one that can have systematic effects on perceptions of stylized facts. We can illustrate this with a specific example. Sub-Saharan Africa has many countries that have small populations, while India and China combined account for about 40% of the world's population. In a decade where India and China did relatively well, such as the 1990s, a country-based analysis will understate the overall improvement in living standards. In contrast, in a decade where Africa did relatively well, such as the 1960s, the overall growth record would appear less strong if assessed on a population-weighted basis. The point that countries differ greatly in terms of population size is important when interpreting tables, graphs and regressions that use the country as the unit of observation.

[7]The figures we use for participation rates are those implicit in the Penn World Table, 6.1.

iii. differences in levels of GDP per worker

Initially, we document the international disparities in GDP per worker. We first look at data for countries with large populations. Table 1 lists a set of countries that together account for 4.3 billion people. Of the countries with large populations, the main omissions are Germany, because of the difficulty posed by reunification, and economies that were centrally planned, including Russia.

Table 1: International Disparities in GDP per Worker

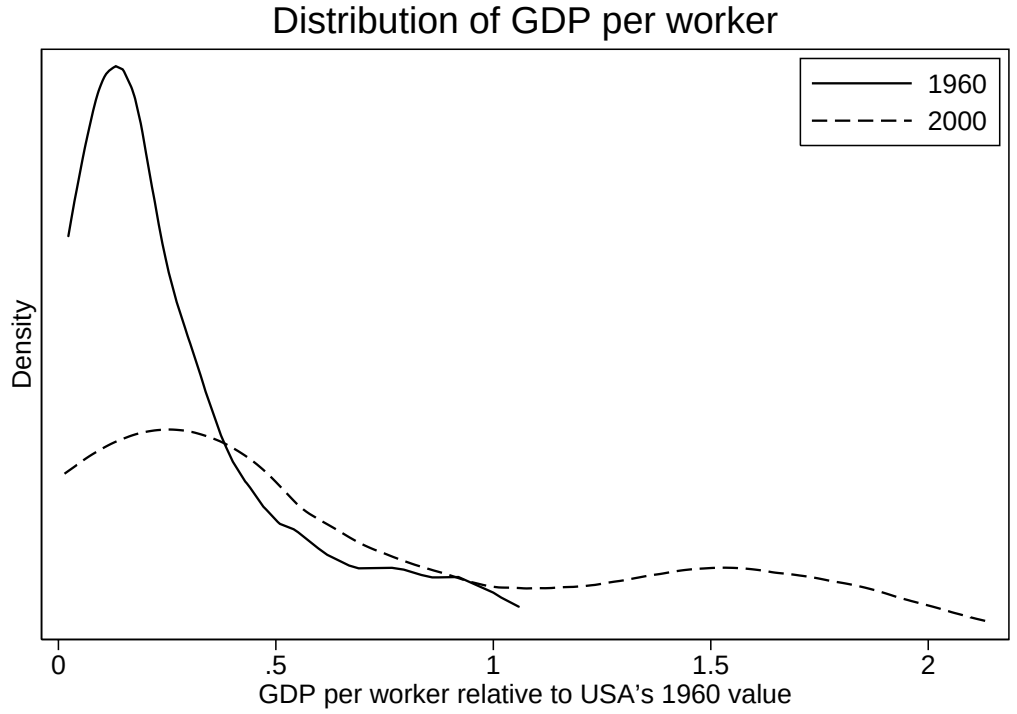
Country	Population(m, 2000)	R1960	R2000
USA	275	1	1
United Kingdom	60	.69	.69
Argentina	37	.62	.40
France	60	.60	.76
Italy	58	.55	.84
South Africa	43	.47	.34
Mexico	97	.44	.38
Spain	40	.40	.68
Iran	64	.30	.30
Colombia	42	.27	.18
Japan	127	.25	.60
Brazil	170	.24	.30
Turkey	67	.17	.24
Philippines	76	.17	.13
Egypt	64	.17	.21
Korea, Republic of	47	.15	.57
Bangladesh	131	.10	.10
Nigeria	127	.08	.02
Indonesia	210	.08	.14
Thailand	61	.07	.20
Pakistan	138	.07	.11
India	1016	.06	.10
China	1259	.04	.10
Ethiopia	64	.04	.02
Mean		.29	.35
Median		.21	.27

Notes:
- R is GDP per worker as a fraction of that in the USA.

The table shows GDP per worker, relative to the USA, for 1960 and 2000. The countries are ranked in descending order in terms of their 1960 position. Some clear patterns emerge: the major economies of Western Europe have maintained their position relative to the USA (as in the case of the UK) or substantially improved it (France, Italy, Spain). Among the poorer nations, there are some countries that have improved their relative position dramatically (Japan, Republic of Korea, Thailand) and others that have performed badly (Argentina, Nigeria). If we look at the mean and median of relative

GDP per worker, there has been a moderate increase, suggesting a slight tendency for reduced dispersion. But these statistics disguise a wide variety of experience, and we will discuss the issue of convergence in more detail below.

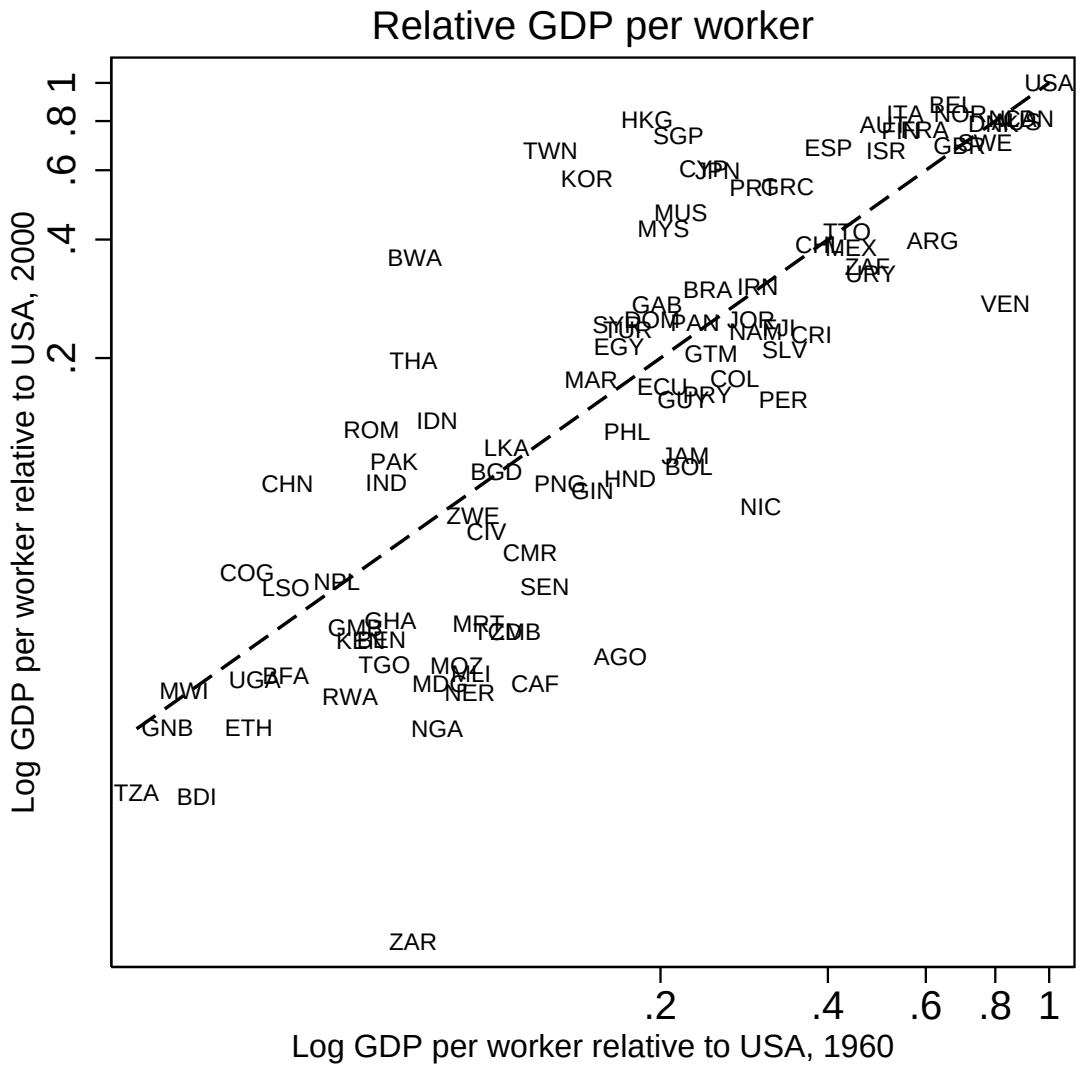
Figure 1: Cross-Country Density of Output per Worker



We now consider the shape of the international distribution of GDP per worker, using the USA's 1960 value as the benchmark. Figure 1 shows a kernel density plot of the distribution of GDP per worker in 1960 and 2000, relative to the benchmark. The rightwards movement reflects the growth that took place over this period. Also noticeable is a thinning in the middle of the distribution, the "Twin Peaks" phenomenon identified in a series of papers by Quah (1993a,b,1996a,b,c,1997).

Is the position in the league table of GDP per worker in 1960 a good predictor of that in 2000? The answer is a qualified yes: the Spearman rank correlation is 0.84. This pattern is shown in more detail in Figure 2, which plots the log of GDP per worker relative to the USA in 2000, against that in 1960. In this and later figures, one or two outlying observations are omitted to facilitate graphing.

Figure 2: Output Per Worker: 1960 versus 2000



The high rank correlation is not a new phenomenon. Easterly et al (1993) report that, for 28 countries for which Maddison (1989) has data, the rank correlation of GDP per capita in 1988 with that in 1870 is 0.82.

iv. growth miracles and disasters

Despite some stability in relative positions, it is easy to pick out countries that have done exceptionally well and others that have done badly. There is an enormous range in observed growth rates, to an extent that has not previously been observed in world history. To show this, we rank the countries by their annual growth rate between 1960 and 2000, and present a list of the fifteen best performers (Table 2) and the fifteen worst (Table 3). To show the dramatic effects of sustaining a high growth rate over forty years, we also show the ratio of GDP per worker in 2000 to that in 1960.

These tables of growth miracles and disasters show a regional pattern that is familiar to anyone who has studied recent economic growth. The best performing countries are mainly located in East Asia and Southeast Asia. These countries have sustained exceptionally high growth rates; for example, GDP per worker has grown by a factor of 11 in the case of Taiwan. If we now turn to the growth disasters, we can see many instances of "negative growth", and these are predominantly countries in sub-Saharan Africa. Later in this section, we will compare Africa's performance with that of other regions in more detail.[8]

[8]Easterly and Levine (1997) and Collier and Gunning (1999a,b) examine various explanations for slow growth in Africa.

Table 2: Fifteen Growth Miracles, 1960-2000

Country	Growth 1960-2000	Factor increase
Taiwan	6.25	11.3
Botswana	6.07	10.6
Hong Kong	5.67	9.09
Korea, Republic of	5.41	8.24
Singapore	5.09	7.29
Thailand	4.50	5.83
Cyprus	4.30	5.39
Japan	4.13	5.04
Ireland	4.10	5.00
China	3.99	4.77
Romania	3.91	4.63
Mauritius	3.88	4.58
Malaysia	3.82	4.48
Portugal	3.48	3.93
Indonesia	3.34	3.72

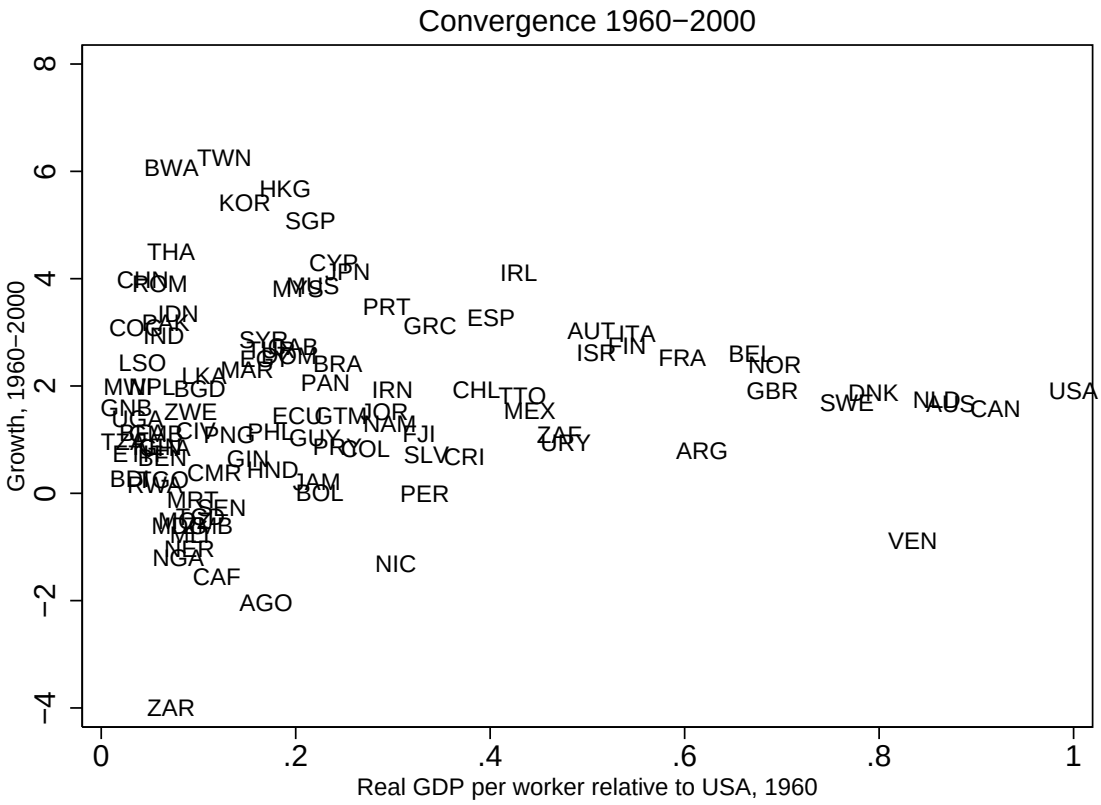
Table 3: Fifteen Growth Disasters, 1960-2000

Country	Growth 1960-2000	Ratio
Peru	0.00	1.00
Mauritania	-0.11	0.96
Senegal	-0.26	0.90
Chad	-0.43	0.84
Mozambique	-0.50	0.82
Madagascar	-0.60	0.79
Zambia	-0.61	0.78
Mali	-0.77	0.74
Venezuela	-0.88	0.70
Niger	-1.03	0.66
Nigeria	-1.21	0.62
Nicaragua	-1.30	0.59
Central African Republic	-1.56	0.53
Angola	-2.04	0.44
Congo, Democratic Rep.	-4.00	0.20

v. convergence?

An alternative way of showing the diversity of experience is to plot the growth rate over 1960-2000 against the 1960 level of real GDP per worker, relative to the USA. This is shown in Figure 3. The most obvious lesson to be drawn from this figure is the diversity of growth rates, especially at low levels of development. The figure does not provide much support for the idea that countries are converging to a common level of income, since that would require evidence of a downward sloping relationship between growth and initial income. Neither does it support the widespread idea that poorer countries have always grown slowly.

Figure 3: Growth Versus Initial Income: 1960-2000



vi. the growth slowdown

Next, we present similar figures for two sub-periods, 1960-1980 and 1980-2000. These plots, shown as Figures 4 and 5, reveal another important pattern. For many developing countries, growth was significantly lower in the second period, with many countries seeing a decline in real GDP per worker after 1980. We can see this more clearly by looking at the international distribution of growth rates for the two sub-periods. Figure 6 shows kernel density estimates, and reveals a clear pattern: the mass of the distribution has shifted leftwards (slower growth) while at the same time the variance has increased (greater dispersion in growth rates).

Figure 4: Growth Versus Initial Income 1960-1980

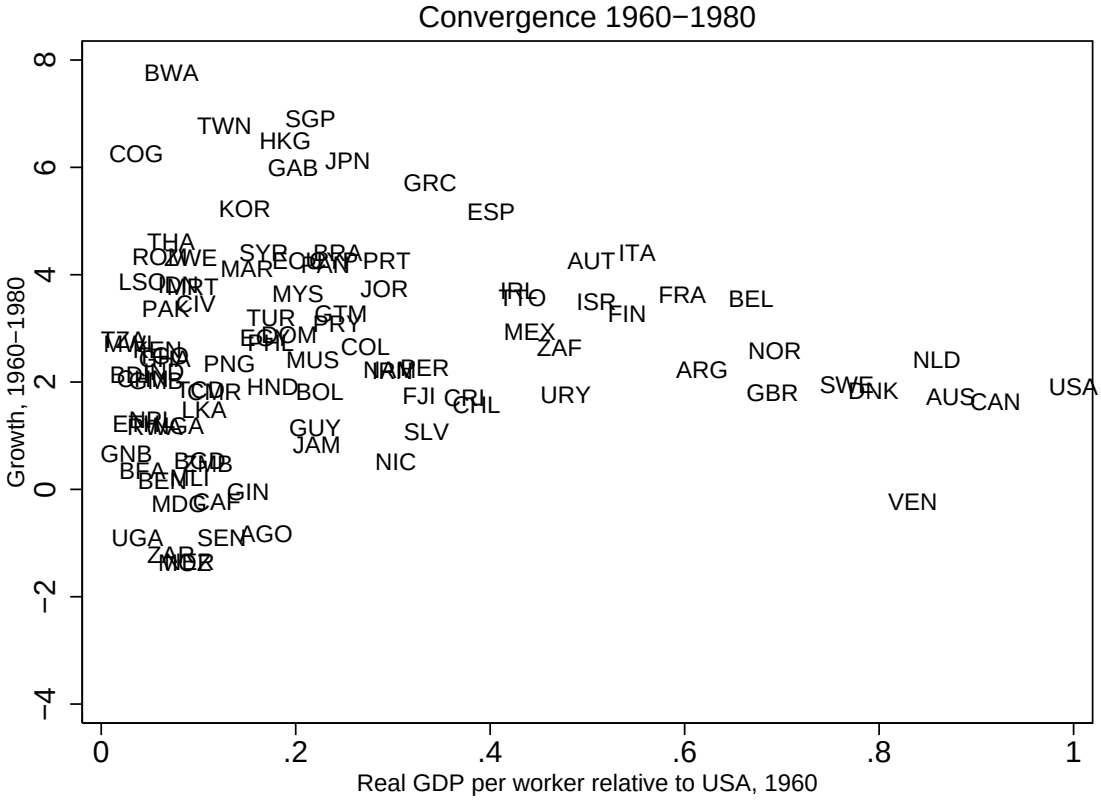


Figure 5: Growth Versus Initial Income: 1980-2000

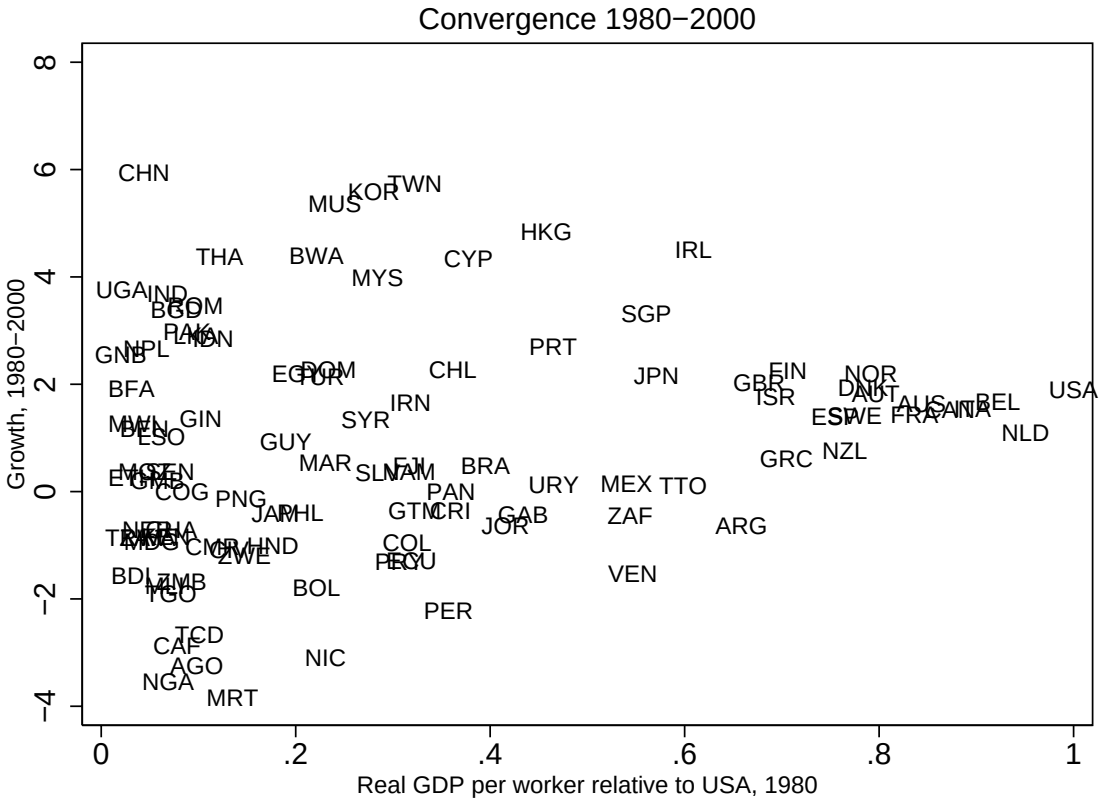


Figure 6: Density of Growth Rates across Countries

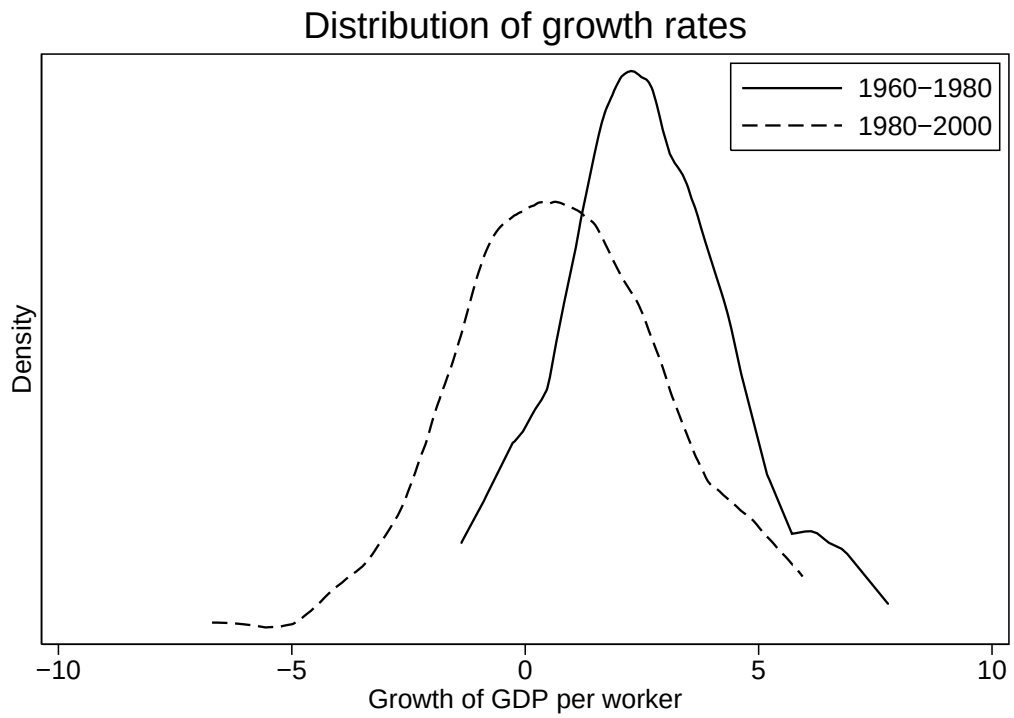
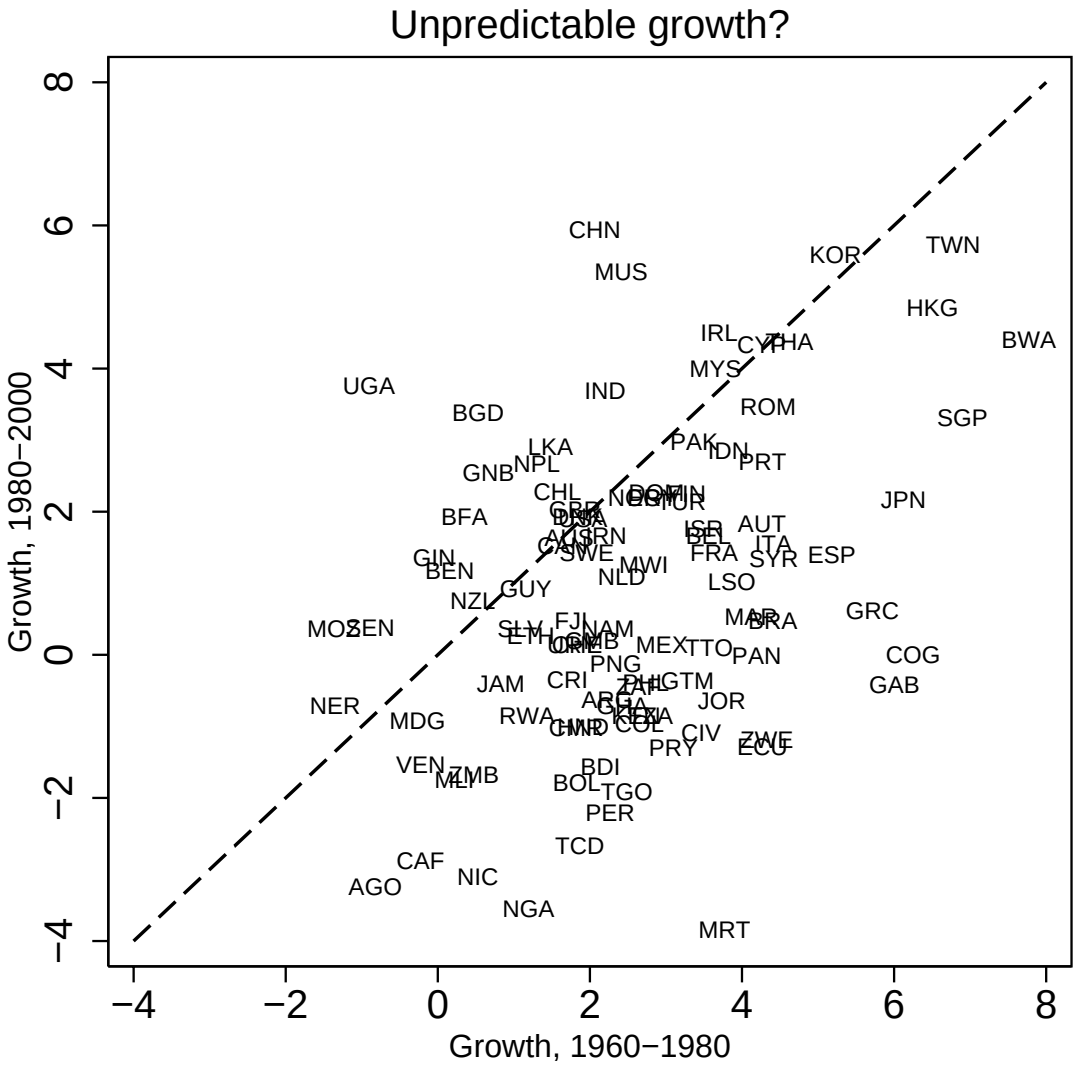


Figure 7: Growth Rates in 1960-1980 versus 1980-2000



A different way to highlight the growth slowdown is to plot the growth rate in 1980-2000 against that in 1960-1980 as is done in Figure 7, which also includes a 45 degree line. Countries above the line have seen growth increase, whereas countries below have seen growth decline. There are clearly more countries in which growth has declined over time, with the crucial exceptions of China and India, which have seen a dramatic improvement. To reveal the same pattern, Table 4 lists the countries in various categories, classified by growth rates in 1960-80 and in 1980-2000.

Table 4: Growth in 1960-1980 and 1980-2000

	G2≤0	0<G2≤1.5	1.5<G2≤3	G2>3
G1<=0	Angola, Central African Republic, DR Congo, Madagascar, Niger, Venezuela	Guinea, Mozambique, Senegal		Uganda
0<G1≤1.5	Jamaica, Mali, Nicaragua, Nigeria, Rwanda, Zambia	Benin, El Salvador, Ethiopia, Guyana, New Zealand	Burkina Faso, Guinea-Bissau, Nepal, Sri Lanka	Bangladesh
1.5<G1≤3	Argentina, Bolivia, Burundi, Cameroon, Chad, Colombia, Costa Rica, Ghana, Honduras, Kenya, Papua New Guinea, Peru, Philippines, South Africa, Tanzania, Togo	Fiji, Gambia, Malawi, Mexico, Namibia, Netherlands, Sweden, Switzerland, Uruguay	Australia, Canada, Denmark, Chile, Dominican Rep., Egypt, Iran, Norway, UK, USA	China, India, Mauritius
G1>3	Ecuador, Gabon, Guatemala, Ivory Coast, Jordan, Mauritania, Panama, Paraguay, Zimbabwe	Brazil, Rep. Congo, France, Greece, Lesotho, Morocco, Spain, Syria, Trinidad and Tobago	Austria, Belgium, Finland, Indonesia, Israel, Italy, Japan, Pakistan, Portugal, Turkey	Botswana, Cyprus, Hong Kong, Ireland, Korea, Malaysia, Romania, Singapore, Taiwan, Thailand

Notes:

- The above table classifies countries according to their annual growth rates over 1960-80 (G1) and over 1980-2000 (G2).

vii. does past growth predict future growth?

Another lesson to be drawn from Figure 7 and Table 4 is that relative performance has been unstable. The correlation between growth in 1960-1980 and that in 1980-2000 is just 0.40, so past growth is not a particularly useful predictor of future growth.[9] For the whole sample, the correlations across decades are also weak (Table 5). It is less well known that the cross-decade correlation has tended to increase over time, as is clear from Table 5's below diagonal elements for the whole sample. This is tentative evidence that national economies are gradually sorting themselves into a pattern of distinct winners and losers.

Table 5: Growth Rate Correlations Across Decades

	1960-1970	1970-1980	1980-1990	1990-2000
Whole sample				
Growth 1960-1970	1.00			
Growth 1970-1980	0.16	1.00		
Growth 1980-1990	0.28	0.31	1.00	
Growth 1990-2000	0.11	0.33	0.44	1.00
Rich country group				
Growth 1960-1970	1.00			
Growth 1970-1980	0.73	1.00		
Growth 1980-1990	0.06	0.40	1.00	
Growth 1990-2000	-0.07	0.37	0.61	1.00

Notes:
- Whole sample is 102 countries. Rich country group is 19 countries.

[9]Easterly et al (1993) emphasized this point, and suggested that the lack of persistence in growth rates indicates the importance of good luck.

viii. Growth differences by development level and geographic region

Can we say anything more about the characteristics of the winners and losers? First, we investigate the relationship between growth and initial development levels in more detail. We rank the sample of 102 countries by initial income in 1960, and then look at the distribution of growth rates for subgroups. In Table 6, for various ranges of initial income relative to the USA, we show the growth rate at the 25th percentile, the median, and the 75th percentile. If we take the 22 countries which began somewhere between 5% and 10% of GDP per worker in the USA, the annual growth rate at the 25th percentile is negative, but is 2.9% at the 75th percentile. This diversity of experience extends throughout the distribution of relative incomes, but is less pronounced for the richest group.

Table 6: Growth, 1960-2000, by Initial Relative Income

Percentile	N	25th	Median	75th
All	102	0.7	1.6	2.7
Relative income:				
R≤0.05	10	1.0	1.5	2.4
R>0.05 & R≤0.10	22	-0.5	0.9	2.9
R>0.10 & R≤0.25	33	0.4	1.9	2.7
R>0.25 & R≤0.50	19	0.8	1.5	3.1
R>0.50	18	1.6	1.9	2.6

Notes:
- This table shows the 25th, 50th and 75th percentiles of the distribution of growth rates for countries at various levels of development in 1960.
- "R" is GDP per worker in 1960 relative to the US level.

Table 7: Growth, 1960-2000, by Country Groups

Group	N	25th	Median	75th
Sub-Saharan Africa	36	-0.5	0.7	1.3
South and Central America	21	0.4	0.9	1.5
East and Southeast Asia	10	3.8	4.3	5.4
South Asia	7	1.9	2.2	2.9
Industrialized countries	19	1.7	2.4	3.0

Notes:
- This table shows the 25th, 50th and 75th percentiles of the distribution of growth rates for various groups of countries.

Table 7 shows the quartiles of growth rates for countries in different regions.[10] Once again, sub-Saharan Africa is revealed as a weak performer. Within sub-Saharan Africa, even the country at the 75th percentile shows growth of just 1.3%. Performance is slightly better for South and Central America, but still not strong. Against this background, the record of East and Southeast Asia looks all the more remarkable.

[10]These country groupings are not exhaustive; for example Fiji and Papua New Guinea do not appear in any of these groups. Analysis of the group of industrialized countries is subject to the sample selection issue highlighted by DeLong (1988).

In further work (not shown) we have constructed versions of Tables 6 and 7 for 1960-1980 and 1980-2000. These reinforce the patterns already discussed: dispersion of growth rates at all levels of development, major differences across regional groups, and a collapse in growth rates after 1980. Even for the developed countries, growth rates were noticeably lower after 1980 than before, reflecting the well-known productivity slowdown and the reduced potential for catch-up by previously fast-growing countries, such as France, Italy and Japan.

ix. stagnation and output volatility

Some countries did not record fast growth even in the boom of the 1960s. Some have simply stagnated or declined, never sustaining a high or even moderate growth rate for the length of time needed to raise output appreciably. In our sample, there are nine countries that have never exceeded their 1960 level of GDP per worker by more than 30%. Even more striking, a quarter of the countries (26 of 102) never exceeded their 1960 level by more than 60%. To put this in context, a country that grew at an average rate of 2% a year over a forty-year period would see GDP per worker rise by around 120%. Easterly (1994) drew attention to the international prevalence of stagnation, and the failure of some poorer countries to break out of low levels of development.

There are other ways in which the behavior of the poorer countries looks very different to that of rich countries. As emphasized by Pritchett (2000a), it is not uncommon for output to undergo a major collapse in less developed countries (LDCs). To show this, we calculate the largest percentage drop in output over three years recorded for each country, using data from 1960 to the latest available year. The precise statistic we calculate is:

$$100 * \left(1 - \min \left(\frac{Y_{1963}}{Y_{1960}}, \frac{Y_{1964}}{Y_{1961}}, \ldots, \frac{Y_{2000}}{Y_{1997}} \right) \right)$$

The largest ten output falls are shown in Table 8, which shows how dramatic an output collapse can be. Several of these output collapses are associated with periods of

intense civil war, as in the cases of Rwanda, Angola and the Democratic Republic of the Congo. But the phenomenon of output collapse is a great deal more widespread than may be explained by events of this type. Of the 102 countries in our sample, 50 showed at least one three-year output collapse of 15% or more. 65 countries experienced a three-year output collapse of 10% or more. In contrast, between 1960 and 2000, the largest three-year output collapse in the USA was 5.4%, and in the UK 3.6%, both recorded in 1979-82. A corollary of these patterns is that time series modeling of LDC output, whether on a country-by-country basis or using panel data, has to be approached with care. It is not clear that the dynamics of output in the wake of a major collapse would look anything like the dynamics at other times.

Table 8: Output Collapses

Country	Largest 3-year drop	Dates
Chad	50%	1980-83
Rwanda	47%	1991-94
Angola	46%	1973-76
Romania	37%	1977-80
Dem. Rep. Congo	36%	1992-95
Mauritania	34%	1985-88
Tanzania	34%	1987-90
Mali	34%	1985-88
Cameroon	33%	1987-90
Nigeria	32%	1997-00

Notes:
- This table shows the ten countries with the largest output collapses over a three-year period, using data on GDP per worker between 1960 and the latest available year.

We conclude our consideration of stylized facts by briefly reporting some evidence on long-run output volatility. Table 9 reports figures on the standard deviation of annual growth rates between 1960 and 2000. Industrialized countries are relatively stable, while sub-Saharan Africa is by far the most volatile region, followed by South and Central America. Volatility is not uniformly higher in developing countries, however: using the standard deviation of annual growth rates, South Africa is less volatile than the USA, Sri Lanka less volatile than Canada, and Pakistan less volatile than Switzerland.

Table 9: Volatility, 1960-2000, by Regions

Group	N	25th	Median	75th
Sub-Saharan Africa	36	5.5	7.4	9.3
South and Central America	21	3.9	4.8	5.4
East and Southeast Asia	10	3.8	4.1	4.7
South Asia	7	3.0	3.3	5.2
Industrialized countries	19	2.3	2.9	3.5

Notes:

- This table shows the 25th, 50th and 75th percentiles of the distribution of the standard deviation of annual growth rates, using data from the earliest available year until the latest available, between 1960 and 2000.

x. a summary of the stylized facts

The stylized facts we consider can be summarized as follows:

1. Over the forty-year period as a whole, most countries have grown richer, but vast income disparities remain. For all but the richest group, growth rates have differed to an unprecedented extent, regardless of the initial level of development.

2. Although past growth is a surprisingly weak predictor of future growth, it is slowly becoming more accurate over time, and so distinct winners and losers are beginning to emerge. The strongest performers are located in East and Southeast Asia, which have sustained growth rates at unprecedented levels. The weakest performers are predominantly located in sub-Saharan Africa, where some countries have barely grown at all, or even become poorer. The record in South and Central America is also distinctly mixed. In these regions, output volatility is high, and dramatic output collapses are not uncommon.

3. For many countries, growth rates were lower in 1980-2000 than in 1960-1980, and this growth slowdown is observed throughout most of the income distribution. Moreover, the dispersion of growth rates has increased. A more optimistic reading would also

emphasize the growth take-off that has taken place in China and India, home to two-fifths of the world's population and a greater proportion of the world's poor.

Even this brief overview of the stylized facts reveals that there is much of interest to be investigated and understood. The field of growth econometrics has emerged through efforts to interpret and understand these facts in terms of simple statistical models, and in the light of predictions made by particular theoretical structures. In either case, the complexity of the growth process and the paucity of the available data combine to suggest that scientific standards of proof are unattainable. Perhaps the best this literature can hope for is to constrain what can legitimately be claimed.

Researchers such as Levine and Renelt (1991) and Wacziarg (2002) have argued that, seen in this more modest light, growth econometrics can provide a signpost to interesting patterns and partial correlations, and even rule out some versions of the world that might otherwise seem plausible. Seen in terms of establishing stylized facts, empirical studies help to broaden the demands made of future theories, and can act as a discipline on quantitative investigations using calibrated models. In the remainder of this chapter, we will discuss in more detail the uses and limits of statistical evidence. We first examine how empirical growth studies are related to theoretical models, and then return in more depth to the study of convergence.

III. Cross-country growth regressions: from theory to empirics

The stylized facts of economic growth have led to two major themes in the development of formal econometric analyses of growth. The first theme revolves around the question of convergence: are contemporary differences in aggregate economies transient over sufficiently long time horizons? The second theme concerns the identification of growth determinants: which factors seem to explain observed differences in growth? These questions are closely related in that each requires the specification of a statistical model of cross-country growth differences from which the effects on growth of various factors, including initial conditions, may be identified. In this section, we

describe how statistical models of cross-country growth differences have been derived from theoretical growth models.

Section III.i provides a general theoretical framework for understanding growth dynamics. The framework is explicitly neoclassical and represents the basis for most empirical growth work; even those studies that have attempted to produce evidence in favor of endogenous or other alternative growth theories have generally used the neoclassical model as a baseline from which to explore deviations. Section III.ii examines the relationship between this theoretical model of growth dynamics and the specification of a growth regression. This transition from theory to econometrics produces the canonical cross-country growth regression.

i. growth dynamics: basic ideas

For economy *i* at time *t*, let $Y_{i,t}$ denote output, $L_{i,t}$ the labor force assumed to obey $L_{i,t} = L_{i,0}e^{n_i t}$ where the population growth rate n_i is constant, and $A_{i,t}$ the efficiency level of each worker with $A_{i,t} = A_{i,0}e^{g_i t}$ where g_i is the (constant) rate of (labor augmenting) technological progress. We will work with two main per capita notions: output per efficiency unit of labor input, $y^E_{i,t} = \frac{Y_{i,t}}{A_{i,t}L_{i,t}}$ and output per labor unit $y_{i,t} = \frac{Y_{i,t}}{L_{i,t}}$. As is well known, the generic one-sector growth model, in either its Solow-Swan or Ramsey-Cass-Koopmans variant, implies, to a first-order approximation, that

$$\log y^E_{i,t} = \left(1 - e^{-\lambda_i t}\right) \log y^E_{i,\infty} + e^{-\lambda_i t} \log y^E_{i,0} \qquad (1)$$

where $y^E_{i,\infty}$ is the steady-state value of $y^E_{i,t}$ and $\lim_{t\to\infty} y^E_{i,t} = y^E_{i,\infty}$. The parameter λ_i (which must be positive) measures the rate of convergence of $y^E_{i,t}$ to its steady-state value

and depends on the other parameters of the model. Given $\lambda_i > 0$, the value of $y_{i,\infty}^E$ is independent of $y_{i,0}^E$ so that, in this sense, initial conditions do not matter in the long-run.[11]

Eq. (1) expresses growth dynamics in terms of the unobservable $y_{i,t}^E$. In order to describe dynamics in terms of the observable variable $y_{i,t}$ we can write equation (1) as

$$\log y_{i,t} - g_i t - \log A_{i,0} = \left(1 - e^{-\lambda_i t}\right) \log y_{i,\infty}^E + e^{-\lambda_i t} \left(\log y_{i,0} - \log A_{i,0}\right) \tag{2}$$

so that

$$\log y_{i,t} = g_i t + \left(1 - e^{-\lambda_i t}\right) \log y_{i,\infty}^E + \left(1 - e^{-\lambda_i t}\right) \log A_{i,0} + e^{-\lambda_i t} \log y_{i,0} \tag{3}$$

In parallel to equation (1), one can easily see that

$$\lim_{t \to \infty} \left(y_{i,t} - y_{i,\infty}^E A_{i,0} e^{g_i t}\right) = 0 \tag{4}$$

so that the initial value of output per worker has no implications for its long-run value.

This description of the dynamics of output provides the basis for describing the dynamics of growth. Let

$$\gamma_i = t^{-1} \left(\log y_{i,t} - \log y_{i,0}\right) \tag{5}$$

denote the growth rate of output per worker between 0 and *t*. Subtracting $\log y_{i,0}$ from both sides of (3) and dividing by *t* yields

$$\gamma_i = g_i + \beta_i \left(\log y_{i,0} - \log y_{i,\infty}^E - \log A_{i,0}\right) \tag{6}$$

[11]Implicit in our discussion is the assumption that $y_{i,0}^E > 0$ which eliminates the trivial equilibrium $y_{i,t}^E = 0 \ \forall t$.

where

$$\beta_i = -t^{-1}\left(1 - e^{-\lambda_i t}\right) \tag{7}$$

The β_i parameter will prove to play a key role in empirical growth analysis.

Equation (6) thus decomposes the growth rate in country i into two distinct components. The first component, g_i, measures growth due to technological progress, whereas the second component $\beta_i\left(\log y_{i,0} - \log y^E_{i,\infty} - \log A_{i,0}\right)$ measures growth due to the gap between initial output per worker and the steady-state value, both measured in terms of efficiency units of labor. This second source of growth is what is meant by "catching up" in the literature. As $t \to \infty$ the importance of the catch-up term, which reflects the role of initial conditions, diminishes to zero.

Under the additional assumptions that the rates of technological progress, and the λ_i parameters are constant across countries, i.e. $g_i = g$, and $\lambda_i = \lambda \ \forall i$, (6) may be rewritten as

$$\gamma_i = g - \beta \log y^E_{i,\infty} - \beta \log A_{i,0} + \beta \log y_{i,0} \tag{8}$$

The important empirical implication of equation (8) is that, in a cross-section of countries, we should observe a negative relationship between average rates of growth and initial levels of output over any time period – countries that start out below their balanced growth path must grow relatively quickly if they are to catch up with other countries that have the same levels of steady-state output per effective worker and initial efficiency. This is closely related to the hypothesis of conditional convergence, which is often understood to mean that countries converge to parallel growth paths, the levels of which are assumed to be a function of a small set of variables.[12] Note, however, that a negative coefficient on initial income in a cross-country growth regression does not automatically imply conditional convergence in this sense, because countries might instead simply be moving toward their own different steady-state growth paths.

[12]We provide formal definitions of convergence in Section IV.i.

ii. cross-country growth regressions

Equation (8) provides the motivation for the standard cross-country growth regression that is the foundation of the empirical growth literature. Typically, these regression specifications start with (8) and append a random error term υ_i so that

$$\gamma_i = g - \beta \log y_{i,\infty}^E - \beta \log A_{i,0} + \beta \log y_{i,0} + \upsilon_i \tag{9}$$

Implementation of (9) requires the development of empirical analogs for $\log y_{i,\infty}^E$ and $\log A_{i,0}$. Mankiw, Romer, and Weil (1992) in a pioneering analysis, show how to do this in a way that produces a growth regression model that is linear in observable variables. In their analysis, aggregate output is assumed to obey a three-factor Cobb-Douglas production function

$$Y_{i,t} = K_{i,t}^{\alpha} H_{i,t}^{\phi} \left(A_{i,t} L_{i,t} \right)^{1-\alpha-\phi} \tag{10}$$

where $K_{i,t}$ denotes physical capital and $H_{i,t}$ denotes human capital. Physical and human capital are assumed to follow the continuous time accumulation equations

$$\dot{K}_{i,t} = s_{K,i} Y_{i,t} - \delta K_{i,t} \tag{11}$$

and

$$\dot{H}_{i,t} = s_{H,i} Y_{i,t} - \delta H_{i,t} \tag{12}$$

respectively, where δ denotes the depreciation rate, $s_{K,i}$ is the saving rate for physical capital and $s_{H,i}$ is the saving rate for human capital and dots above variables denote time

derivatives. Note that the saving rates are both assumed to be time invariant. These accumulation equations, combined with the parameter constancy assumptions used to justify eq. (8) imply that the steady-state value of output per effective worker is

$$y_{i,\infty}^{E} = \left(\frac{s_{K,i}^{\alpha} s_{H,i}^{\phi}}{\left(n_i + g + \delta\right)^{\alpha+\phi}} \right)^{\frac{1}{1-\alpha-\phi}} \tag{13}$$

producing a cross-country growth regression of the form

$$\gamma_i = g + \beta \log y_{i,0} + \beta \frac{\alpha+\phi}{1-\alpha-\phi} \log\left(n_i + g + \delta\right) - \beta \frac{\alpha}{1-\alpha-\phi} \log s_{K,i} - \beta \frac{\phi}{1-\alpha-\phi} \log s_{H,i} - \beta \log A_{i,0} + \upsilon_i \tag{14}$$

Mankiw, Romer and Weil assume that $A_{i,0}$ is unobservable and that $g + \delta$ is known. These assumptions mean that (14) is linear in the logs of various observable variables and therefore amenable to standard regression analysis.

Mankiw, Romer, and Weil argue that $A_{i,0}$ should be interpreted as reflecting not just technology, which they assume to be constant across countries, but country-specific influences on growth such as resource endowments, climate and institutions. They assume these differences vary randomly in the sense that

$$\log A_{i,0} = \log A + e_i \tag{15}$$

where e_i is a country-specific shock distributed independently of n_i, $s_{K,i}$, and $s_{H,i}$[13]. Substituting this into (14) and defining $\varepsilon_i = \upsilon_i - \beta e_i$, we have the regression relationship

[13]This independence assumption is justified, in turn, on the basis that 1) n_i, $s_{K,i}$, and $s_{H,i}$ are exogenous in the neoclassical model with isoelastic preferences and 2) the estimated parameter values are consistent with those predicted by the model.

$$\gamma_i = g - \beta \log A + \beta \log y_{i,0} + \beta \frac{\alpha+\phi}{1-\alpha-\phi} \log\left(n_i + g + \delta\right) - \beta \frac{\alpha}{1-\alpha-\phi} \log s_{K,i} - \beta \frac{\phi}{1-\alpha-\phi} \log s_{H,i} + \varepsilon_i \quad (16)$$

Using data from a group of 98 countries over the period 1960 to 1985, Mankiw Romer and Weil produce regression estimates of $\hat{\beta} = -.299$, $\hat{\alpha} = .48$ and $\hat{\phi} = .23$.[14] [15] Mankiw, Romer, and Weil are unable to reject the overidentifying restrictions present in (16). While this result is echoed in studies such as Knight, Loayza, and Villenueva (1993), other authors, Caselli, Equivel, and Lefort (1996), for example, are able to reject the restrictions.

Many cross-country regression studies have attempted to extend Mankiw, Romer and Weil by adding additional control variables Z_i to the regression suggested by (16). Relative to Mankiw, Romer and Weil, such studies may be understood as allowing for predictable heterogeneneity in the steady-state growth term g_i and initial technology term $A_{i,0}$ that are assumed constant across *i* in (16). Formally, the $g_i - \beta \log A_{i,0}$ terms in (6) are replaced with $g - \beta \log A + \pi Z_i - \beta e_i$ rather than with $g - \beta \log A - \beta e_i$ which produced (16). (As far as we know, empirical work universally ignores the fact that $\log\left(n_i + g + \delta\right)$ should also be replaced with $\log\left(n_i + g_i + \delta\right)$.) This produces the cross country growth regression

$$\gamma_i = g - \beta \log A + \beta \log y_{i,0} + \beta \frac{\alpha+\phi}{1-\alpha-\phi} \log\left(n_i + g + \delta\right) - \beta \frac{\alpha}{1-\alpha-\phi} \log s_{K,i} - \beta \frac{\phi}{1-\alpha-\phi} \log s_{H,i} + \pi Z_i + \varepsilon_i \quad (17)$$

[14] Based on data from the US and other economies, Mankiw, Romer, and Weil set $g + \delta = .05$ prior to estimation.

[15] Using $\lambda = -t^{-1} \log\left(1 - t\beta\right)$, the implied estimate of λ is .0142. The relationship $\lambda_i = \left(1-\alpha-\phi\right)\left(n_i + g + \delta\right)$ was not imposed by Mankiw, Romer and Weil, who instead treat λ as a constant to be estimated. Durlauf and Johnson (1995, Table II, note b) show that estimating this model when λ varies with *n* in the way implied by the theory produces only very small changes in parameter estimates.

. The regression described by (17) does not identify whether the controls Z_i are correlated with steady-state growth g_i or the initial technology term $A_{i,0}$. For this reason, a believer in a common steady-state growth rate will not be dissuaded by the finding that particular choices of Z_i help predict growth beyond the Solow regressors. Nevertheless, it seems plausible that the controls Z_i may sometimes function as proxies for predicting differences in efficiency growth g_i rather than in the initial technology $A_{i,0}$. As argued in Temple (1999), even if all countries have the same total factor productivity growth (TFP) in the long run, over a twenty- or thirty-year sample the assumption of equal TFP growth is highly implausible, so the variables in Z_i can explain these differences. That being said, the attribution of the predictive content of Z_i to initial technology versus steady state growth will entirely depend on a researcher's prior beliefs. It is possible that proper accounting of the $\log(n_i + g_i + \delta)$ term would allow for some progress in identifying g_i versus $A_{i,0}$ effects since g_i effects would imply a nonlinear relationship between Z_i and overall growth γ_i; however this nonlinearity may be too subtle to uncover given the relatively small data sets available to growth researchers.

The canonical cross-country growth regression may understood as a version of (17) when the cross-coefficient restrictions embedded in (17) are ignored (which is usually the case in empirical work). A generic representation of the regression is

$$\gamma_i = \beta \log y_{i,0} + \psi X_i + \pi Z_i + \varepsilon_i \tag{18}$$

where X_i contains a constant, $\log(n_i + g + \delta)$, $\log s_{K,i}$ and $\log s_{H,i}$. The variables spanned by $\log y_{i,0}$ and X_i thus represent those growth determinants that are suggested by the Solow growth model whereas Z_i represents those growth determinants that lie outside Solow's original theory.[16] The distinction between the Solow variables and Z_i is

[16]We distinguish $\log y_{i,0}$ from the other Solow variables because of the role it plays in analysis of convergence; see Section IV for detailed discussion.

important in understanding the empirical literature. While the Solow variables usually appear in different empirical studies, reflecting the treatment of the Solow model as a baseline for growth analysis, choices concerning which Z_i variables to include vary greatly.

Equation (18) represents the baseline for much of growth econometrics. These regressions are sometimes known as Barro regressions, given Barro's extensive use of such regressions to study alternative growth determinants starting with Barro (1991). This regression model has been the workhorse of empirical growth research.[17] In modern empirical analyses, the equation has been generalized in a number of dimensions. Some of these extensions reflect the application of (18) to time series and panel data settings. Other generalizations have introduced nonlinearities and parameter heterogeneity. We will discuss these variants below.

iii. interpreting errors in growth regressions

Our development of the relationship between cross-country growth regressions and neoclassical growth theories illustrates the standard practice of adding regression errors in an ad hoc fashion. Put differently, researchers usually derive a deterministic growth relationship and append an error in order to capture whatever aspects of the growth process are omitted from the model that has been developed. One problem with this practice is that some types of errors have important implications for the asymptotics of estimators. Binder and Pesaran (1999) conduct an exhaustive study of this question, one important conclusion of which is that if one generalizes the assumption of a constant rate of technical change so that technical change follows a random walk, this induces nonstationarity in many levels series, raising attendant unit root questions.

[17]Such regressions appear to have been employed earlier by Grier and Tullock (1989) and Kormendi and Meguire (1985). The reason these latter two studies seem to have received less attention than warranted by their originality is, we suspect, due to their appearance before endogenous growth theory emerged as a primary area of macroeconomic research, in turn placing great interest on the empirical evaluation of growth theories. To be clear, Barro's development is original to him and his linking of cross-country growth regressions to alternative growth theories was unique.

Beyond issues of asymptotics, the ad hoc treatment of regression errors leaves unanswered the question of what sorts of implicit substantive economic assumptions are made by a researcher who does this. Brock and Durlauf (2001a) address this issue using the concept of exchangeability. Basically, their argument is that in a regression such as (18), a researcher typically thinks of the errors ε_i as interchangeable across observations: different patterns of realized errors are equally likely to occur if the realizations are permuted across countries. In other words, the information available to a researcher about the countries is not informative about the error terms.

Exchangeability is a mathematical formalization of this idea and is defined as follows. For each observation *i*, there exists an associated information set F_i available to the researcher. In the growth context, F_i may include knowledge of a country's history or culture as well as any "economic" variables that are known. A definition of exchangeability (formally, *F*-conditional exchangeability) is

$$\mu\left(\varepsilon_1 = a_1, ..., \varepsilon_N = a_N \left| F_1 ... F_N \right.\right) = \mu\left(\varepsilon_{\rho(1)} = a_1, ..., \varepsilon_{\rho(N)} = a_N \left| F_1 ... F_N \right.\right) \qquad (19)$$

where $\rho(\ )$ is an operator that permutes the *N* indices.

Many criticisms of growth regressions amount to arguments that exchangeability has been violated. For example, omitted regressors induce exchangeability violations as these regressors are elements of F. Parameter heterogeneity also leads to nonexchangeability. For these cases, the failure of nonexchangeability calls into question the interpretation of the regression. This is not always the case; heteroskedasticity in errors violates exchangeability but does not induce interpretation problems for coefficients.

Brock and Durlauf argue that exchangeability produces a link between substantive social science knowledge and error structure, i.e. this knowledge may be used to evaluate the plausibility of exchangeability. They suggest that a good empirical practice would for researchers to question whether the errors in a model are exchangeable, and if not, determine whether the violation invalidates the purposes for which the regression is being used. This cannot be done in an algorithmic fashion, but as is the case with empirical

work quite generally, requires judgments by the analyst. See Draper et al (1993) for further discussion of the role of exchangeability in empirical work.

IV. The convergence hypothesis

Much of the empirical growth literature has focused on the convergence hypothesis. Although questions of convergence predate them, recent widespread interest in the convergence hypothesis originates from Abramovitz (1986) and Baumol (1986). This interest and the availability of the requisite data for a broad cross-section of countries, due to Summers and Heston (1988,1991), spawned an enormous literature testing the convergence hypothesis in one or more of its various guises.[18]

In this section, we explore the convergence hypothesis. In Section IV.i we consider the specification of notions of convergence as related to the relationship between initial conditions and long-run outcomes. Section IV.ii explores the main technique that has been employed in studying long-run dependence, β-convergence. Section IV.iii considers alternative notions of convergence that focus less on the persistence of initial conditions and instead on whether the cross-section dispersion of incomes is decreasing across time. This section explores both σ-convergence and more general notions and recent methods that fall under the heading of distributional dynamics. It also considers how distibutional notions of convergence may be related to definitions found in Section IV.i. Section IV.iv develops time series approaches to convergence. Section IV.v moves beyond the question of whether convergence is present to consider analyses that have attempted to identify the sources of convergence when it appears to be present.

i. **convergence and initial conditions**

The effect of initial conditions on long-run outcomes arguably represents the primary empirical question that has been explored by growth economists. The claim that

[18]See Durlauf (1996) and the subsequent papers in the July 1996 *Economic Journal*, Durlauf and Quah (1999), Islam (2003) and Barro and Sala-i-Martin (2004) for surveys of aspects of the convergence literature.

the effects of initial conditions eventually disappear is the heuristic basis for what is known as the convergence hypothesis. The goal of this literature is to answer two questions concerning per capita income differences across countries (or other economic units, such as regions). First, are the observed cross-country differences in per capita incomes temporary or permanent? Second, if they are permanent, does that permanence reflect structural heterogeneity or the role of initial conditions in determining long-run outcomes? If the differences in per capita incomes are temporary, unconditional convergence (to a common long-run level) is occurring. If the differences are permanent solely because of cross-country structural heterogeneity, conditional convergence is occurring. If initial conditions determine, in part at least, long-run outcomes, and countries with similar initial conditions exhibit similar long-run outcomes, then one can speak of convergence clubs.[19]

We first consider how to formalize the idea that initial conditions matter. While the discussion focuses on $\log y_{i,t}$, the log level of per capita output in country i at time *t*; these definitions can in principle be applied to other variables such as real wages, life expectancy, etc. Our use of $\log y_{i,t}$ rather than $y_{i,t}$ reflects the general interest in the growth literature in relative versus absolute inequality, i.e. one is usually more interested in whether the ratio of income between two countries exhibits persistence than an absolute difference, particularly since sustained economic growth will imply that a constant levels difference is of asymptotically negligible size when relative income is considered.

We associate with $\log y_{i,t}$ initial conditions, $\rho_{i,0}$. These initial conditions do not matter in the long-run if

$$\lim_{t\to\infty} \mu\left(\log y_{i,t} \middle| \rho_{i,0}\right) \textit{does not depend on } \rho_{\mathrm{i},0} \qquad (20)$$

[19]This taxonomy is due to Galor (1996) who discusses the relationship between it and the theoretical growth literature, giving several examples of models in which initial conditions matter for long-run outcomes.

where $\mu(\cdot)$ is a probability measure. To see how this definition connects with empirical growth work, empirical studies of convergence are often focused on whether long-run per capita output depends on initial stocks of human and physical capital.

Economic interest in convergence stems from the question of whether certain initial conditions lead to persistent differences in per capita output between countries (or other economic units). One can thus use (20) to define convergence between two economies. Let $\| \ \|$ denote a metric for computing the distance between probability measures.[20] Then countries *i* and *j* exhibit convergence if

$$\lim_{t\to\infty} \left\| \mu\left(\log y_{i,t} \left| \rho_{i,0}\right.\right) - \mu\left(\log y_{j,t} \left| \rho_{j,0}\right.\right) \right\| = 0 \tag{21}$$

Growth economists are generally interested in average income levels; eq. (21) implies that countries *i* and *j* exhibit convergence in average income levels in the sense that

$$\lim_{t\to\infty} E\left(\log y_{i,t} - \log y_{j,t} \left| \rho_{i,0}, \rho_{j,0}\right.\right) = 0. \tag{22}$$

To the extent one is interested in whether countries exhibit common steady-state growth rates, one can modify (22) to require that the limiting expected difference between $\log y_{i,t}$ and $\log y_{j,t}$ is bounded. One way of doing this is due to Pesaran (2004a) and is discussed below.

These notions of convergence can be relaxed. Bernard and Durlauf (1996) suggest a form of partial convergence that relates to whether contemporaneous income differences are expected to diminish. If $\log y_{i,0} > \log y_{j,0}$, their definition amounts to asking whether

$$E\left(\log y_{i,t} - \log y_{j.t} \left| \rho_{i,0}, \rho_{j,0}\right.\right) < \log y_{i,0} - \log y_{j,0} \tag{23}$$

[20]There is no unique or single generally agreed upon metric for measuring deviations between probability measures.

A number of modifications of these definitions have been proposed. Hall, Robertson, and Wickens (1997) suggest appending a requirement that the variance of output differences diminish to 0 over time, i.e.

$$\lim_{t\to\infty} E\left(\left(\log y_{i,t} - \log y_{j,t}\right)^2 \left| \rho_{i,0}, \rho_{j,0} \right.\right) = 0 \tag{24}$$

so that convergence requires output for a pair of countries to behave similarly in the long-run. In our view, this is an excessively strong requirement since it does not allow one to regard the output series as stochastic in the long-run. Eq. (24) would imply that convergence does not occur if countries are perpetually subjected to distinct business cycle shocks. However, Hall, Robertson and Wickens (1997) do identify a weakness of definition (22), namely the failure to control for long-run deviations whose current direction is not predictable. To see this, suppose that $\log y_{i,t} - \log y_{j,t}$ is a random walk with current value 0. In this case, definition (22) would be fulfilled, although output deviations between countries *i* and *j* will become arbitrarily large at some future date.

In recent work, Pesaran (2004a) has proposed a convergence definition that focuses specifically on the likelihood of large long-run deviations. Specifically, Pesaran defines convergence as

$$\lim_{t\Rightarrow\infty} \text{Prob}\left(\left(\log y_{i,t} - \log y_{j,t}\right)^2 < C^2 \left| \rho_{i,0}, \rho_{j,0} \right.\right) > \pi \tag{25}$$

where *C* denotes a deviation magnitude and π is a tolerance probability. The idea of this definition is to focus convergence analysis on output deviations that are economically important and to allow for some flexibility with respect to the probability with which they occur.

These convergence definitions do not allow for the distinction between the long-run effects of initial conditions and the long-run effects of structural heterogeneity. From the perspective of growth theory, this is a serious limitation. For example, the distinctions between endogenous and neoclassical growth theories focus on the long-run

effects of cross-country differences initial human and physical capital stocks; in contrast, cross-country differences in preferences can have long-term effects under either theory. Hence, in empirical work, it is important to be able to distinguish between initial conditions $\rho_{i,0}$ and structural characteristics $\theta_{i,0}$. Steady state effects of initial conditions imply the existence of convergence clubs whereas steady-state effects of structural characteristics do not. In order to allow for this, one can modify (21) so that

$$\lim_{t\to\infty} \left\| \mu\left(\log y_{i,t} \left| \rho_{i,0}, \theta_{i,0}\right.\right) - \mu\left(\log y_{j,t} \left| \rho_{j,0}, \theta_{j,0}\right.\right) \right\| = 0 \text{ if } \theta_{i,0} = \theta_{j,0} \tag{26}$$

implies that countries *i* and *j* exhibit convergence. The notions of convergence in expected value (eq. (22)) may be modified in this way as well,

$$\lim_{t\to\infty} E\left(\log y_{i,t} - \log y_{j,t} \left| \rho_{i,0}, \theta_{i,0}, \rho_{j,0}, \theta_{j,0}\right.\right) = 0 \text{ if } \theta_{i,0} = \theta_{j,0} \tag{27}$$

as can partial convergence in expected value (eq. (23)) and the other convergence concepts discussed above.

In practice, the distinction between initial conditions and structural heterogeneity generally amounts to treating stocks of initial human and physical capital as the former and other variables as the latter. As such, both the Solow variables *X* and the control variables *Z* that appear in cross-country growth regression cf. (18) are usually interpreted as capturing structural heteogeneity. This practice may be criticized if these variables are themselves endogenously determined by initial conditions, a point that will arise below.

The translation of these ideas into restrictions on growth regressions has led to a range of statistical definitions of convergence which we now examine. Before doing so, we emphasize that none of these statistical definitions is necessarily of intrinsic interest per se; rather each concept is useful only to the extent it elucidates economically interesting notions of convergence such as eq. (20). The failure to distinguish between convergence as an economic concept and convergence as a statistical concept has led to a good deal of confusion in the growth literature.

ii. β-convergence

Statistical analyses of convergence have largely focused on the properties of β in regressions of the form (18). β-convergence, defined as $\beta < 0$ is easy to evaluate because it relies on the properties of a linear regression coefficient. It is also easy to interpret in the context of the Solow growth model, since the finding is consistent with the dynamics of the model. The economic intuition for this is simple. If two countries have common steady-state determinants and are converging to a common balanced growth path, the country that begins with a relatively low level of initial income per capita has a lower capital-labor ratio and hence a higher marginal product of capital; a given rate of investment then translates into relatively fast growth for the poorer country. In turn, β-convergence is commonly interpreted as evidence against endogenous growth models of the type studied by Romer and Lucas, since a number of these models specifically predict that high initial income countries will grow faster than low initial income countries, once differences in saving rates and population growth rates have been accounted for. However, not all endogenous growth models imply an absence of β-convergence and therefore caution must be exercised in drawing inferences about the nature of the growth process from the results of β-convergence tests.[21]

There now exists a large body of studies of β-convergence, studies that are differentiated by country set, time period and choice of control variables. When controls are absent, $\beta < 0$ is known as unconditional β-convergence: conditional β-convergence is said to hold if $\beta < 0$ when controls are present. Interest in unconditional β-convergence, while not predicted by the Solow growth model except when countries have common steady-state output levels, derives from interest in the hypothesis that all countries are converging to the same growth path, which is critical in understanding the

[21]Jones and Manuelli (1990) and Kelly (1992) are early examples of endogenous growth models compatible with β-convergence. Each model produces steady state growth without exogenous technical change yet each implies relatively fast growth for initially capital poor economies.

extent to which current international inequality will persist into the far future.[22] Typically, the unconditional β-convergence hypothesis is supported when applied to data from relatively homogeneous groups of economic units such as the states of the US, the OECD, or the regions of Europe; in contrast there is generally no correlation between initial income and growth for data taken from more heterogeneous groups such as a broad sample of countries of the world.[23]

Many cross-section studies employing the β-convergence approach find estimated convergence rates of about 2% per year.[24] This result is found in data from such diverse entities as the countries of the world (after the addition of conditioning variables), the OECD countries, the US states, the Swedish counties, the Japanese prefectures, the regions of Europe, the Canadian provinces, and the Australian states, among others; it is also found in data sets that range over time periods from the 1860's though the 1990's.[25] Some writings go so far as to give this value a status analogous to a universal constant in physics.[26] In fact, there is some variation in estimated convergence

---

[22]Formally, $\beta-$convergence is an implication of (9) if $\log y_{i,\infty}^{E}$ is assumed constant across countries in addition to the assumption on $\log A_{i,0}$ made in (15).

[23]See Barro and Sala-i-Martin (2004, chapters 11 and 12) for application of β-convergence tests to a variety of data sets. Homogeneity can reflect self-selection as pointed out by DeLong (1988). He argues that Baumol's (1986) conclusion that unconditional β-convergence occurred over 1870-1979 among a set of affluent (in 1979) countries is spurious for this reason.

[24]Panel studies estimates of convergence rates have typically been substantially higher than cross-section estimates. Examples where this is true for regressions that only control for the Solow variables include Islam (1995) and Lee, Pesaran, and Smith (1998). The panel approach has possible interpretation problems which we discuss in Section VI.

[25]For example, Barro and Sala-i-Martin (1991) present results for US states and regions as well as European regions; Barro and Sala-i-Martin (1992) for US states, a group of 98 countries and the OECD; Mankiw, Romer, and Weil (1992) for several large groups of countries; Sala-i-Martin (1996a,b) for US states, Japanese prefectures, European regions, and Canadian provinces; Cashin (1995) for Australian states and New Zealand; Cashin and Sahay (1996) for Indian regions; Persson (1997) for Swedish counties; and, Shioji (2001a) for Japanese prefectures and other geographic units.

[26]An alternative view is expressed by Quah (1996b) who suggests that the 2% finding may be a statistical artifact that arises for reasons unrelated to convergence *per se*. At the most primitive level, like any endogenous variable, the rate of convergence is determined by preferences, technology, and endowments. Operationally, this means that the rate of convergence will depend on model parameters and exogenous variables. For example, as

rates, but the range is relatively small; estimates generally range between 1% and 3%, as noted by Barro and Sala-i-Martin (1992).[27]

Despite the many confirmations of this result now in the literature, the claim of global conditional β-convergence remains controversial; here we review the primary problems with the β-convergence literature.

a. robustness with respect to choice of control variables

In moving from unconditional to conditional β-convergence, complexities arise in terms of the specification of steady-state income. The reason for this is the dependence of the steady-state on Z. Theory is not always a good guide in the choice of elements of *Z*; differences in formulations of equation (18) have led to a "growth regression industry" as researchers have added plausibly relevant variables to the baseline Solow specification. As a result, one can identify variants of (18) where convergence appears to occur as $\hat{\beta} < 0$ as well as variants where divergence occurs, i.e. $\hat{\beta} > 0$.

We discuss issues of uncertainty in the specification of growth regressions below. Here we note here that one class of efforts to address model uncertainty has led to confirmatory evidence of conditional β-convergence. This approach assigns probabilities to alternative formulations of (18) and uses these probabilities to construct statements about β that average across the different models. Doppelhofer, Miller, and

---

stated above, in the augmented Solow model studied by Mankiw, Romer, and Weil (1992), the relationship between the rate of convergence and the parameters of the model is $\lambda_i = (1-\alpha-\phi)(n_i + g + \delta)$. Barro and Sala-i-Martin (2004, p. 111-113) discuss the relationship for the case of the Ramsey-Cass-Koopmans model with an isoelastic utility function and a Cobb-Douglas production function. Given this dependence, the ubiquity of the estimated 2% rate of convergence, taken at face value, appears to suggest a remarkable uniformity of preferences, technologies, and endowments across the economic units studied.

[27]Barro and Sala-i-Martin argue that this variation reflects unobserved heterogeneity in steady-state values with more variation being associated with slower convergence. However, in as much as it is correlated with variables included in the regression equations, unobserved heterogeneity renders the parameter estimators inconsistent, which renders the estimated convergence parameter hard to interpret.

Sala-i-Martin (2004) conclude the posterior probability that initial income is part of the linear growth model is 1.00 with a posterior expected value for β of -0.013; this leads to a point estimate of a convergence rate of 1.3% per annum, which is somewhat lower than the 2% touted in the literature; Fernandez, Ley, and Steel (2001a) also find that the posterior probability that initial income is part of the linear growth model is 1.00, despite using a different set of potential models and different priors on model parameters.[28] We therefore conclude that the evidence for conditional β-convergence appears to be robust with respect to choice of controls.

b. identification and nonlinearity: β – convergence and economic divergence

A second problem with the β-convergence literature is an absence of attention to the relationship between β-convergence and economic convergence as defined by eq. (20) or variations based upon it. Put differently, in the β-convergence literature there is a general failure to develop tests of the convergence hypothesis that discriminate between convergent economic models and a rich enough set of non-converging alternatives.

While $\beta < 0$ is an implication of the Solow growth model and so is an implication of the baseline convergent growth model in the literature, this does not mean that $\beta < 0$ is inconsistent with economically interesting non-converging alternatives. One such example is the model of threshold externalities and growth developed by Azariadis and Drazen (1990). In this model, there is a discontinuity in the aggregate production function for aggregate economies. This discontinuity means that the steady-state behavior of a given economy depends on whether its initial capital stock is above or below this threshold; specifically, this model may exhibit two distinct steady states. (Of course, there can be any number of such thresholds.) An important feature of the Azariadis-Drazen model is that data generated by economies that are described by it can exhibit statistical convergence even when multiple steady states are present.

To illustrate this, we follow an argument in Bernard and Durlauf (1996) based on a simplified growth regression. Suppose that for every country in the sample, the Solow

[28]Fernandez, Ley, and Steel (2001a) do not report a posterior expected value for β.

variables X_i and additional controls Z_i are identical. Suppose as well that there is no technical change or population growth. Following the standard arguments for deriving a cross-country regression specification, the growth regression implied by the Azariadis-Drazen assumption on the aggregate production function is

$$\gamma_i = k + \beta\left(\log y_{i,0} - \log y^*_{l(i)}\right) + \varepsilon_i \tag{28}$$

where $l(i)$ indicates the steady state with which country i is associated and $y^*_{l(i)}$ denotes output per capita in that steady state; all countries associated with the same steady state thus have the same $\log y^*_{l(i)}$ value.

The threshold externality model clearly does not exhibit economic convergence as defined above so long as there are at least two steady states. Yet the data generated by a cross-section of countries exhibiting multiple steady states may exhibit statistical convergence. To see this, notice that for this stylized case, the cross-country growth regression may be written as

$$\gamma_i = k + \beta \log y_{i,0} + \varepsilon_i \tag{29}$$

Since the data under study are generated by (28), this standard regression is misspecified. What happens when (29) is estimated when (28) is the data generating process? Using population moments, the estimated convergence parameter β_{ols} will equal

$$\beta_{ols} = \beta \frac{\operatorname{cov}\left(\left(\log y_{i,0} - \log y^*_{l(i)}\right), \log y_{i,0}\right)}{\operatorname{var}(\log y_{i,0})} = \beta\left(1 - \frac{\operatorname{cov}\left(\log y^*_{l(i)}, \log y_{i,0}\right)}{\operatorname{var}\left(\log y_{i,0}\right)}\right) \tag{30}$$

From the perspective of tests of the convergence hypothesis, the noteworthy feature of (30) is that one cannot determine the sign of β_{ols} a priori as it depends on

$1-\frac{\text{cov}\left(\log y^{*}_{l(i)},\log y_{i,0}\right)}{\text{var}\left(\log y_{i,0}\right)}$, which is a function of the covariance between the initial and steady-state incomes of the countries in the sample. It is easy to see that it is possible for β_{ols} to be negative even when the sample includes countries associated with different steady states. Roughly speaking, one would expect $\beta_{ols}<0$ if low-income countries tend to initially be below their steady states whereas high-income countries tend to start above their steady states. While we do not claim this is necessarily the case empirically, the example does illustrate how statistical convergence (defined as $\beta<0$) may be consistent with economic nonconvergence. Interestingly, it is even possible for the estimated convergence parameter β_{ols} to be smaller (and hence imply more rapid convergence) than the structural parameter β in (28).

Below, we review evidence of multiple steady states in the growth process. At this stage, we would note two things. First, some studies have produced evidence of multiple regimes in the sense that statistical models consistent with multiple steady states appear to better fit the cross-country data than the linear Solow model, e.g. Durlauf and Johnson (1995). Second, other studies have produced evidence of parameter heterogeneity such that β appears to depend nonlinearly on initial conditions so that it is equal to 0 for some countries; Liu and Stengos (1999) find precisely this when they reject the specification of constant β for all countries in favor of a specification in which β depends on initial income. These types of findings imply the compatibility of observed growth patterns with the existence of permanent income differences between economies with identical population growth and savings rates and access to identical technologies.

c. endogeneity

A third criticism that is sometimes made of the empirical convergence literature is based on the failure to account for the endogeneity of the explanatory regressors in growth regressions. One obvious reason why endogeneity may matter concerns the consistency of the regression estimates. This concern has led some authors to propose instrumental variables approaches to estimating β. Barro and Lee (1994) analyze

growth data in the periods 1965 to 1975 and 1975 to 1985 and use 5-year lagged explanatory variables as instruments. Barro and Lee find that the use of instrumental variables has little effect on coefficient estimates. Caselli, Esquivel, and Lefort (1996) employ a generalized method of moments (GMM) estimator to analyze a panel variant of the standard cross-country growth regression; growth in the panel is measured in 5-year intervals for 1960-1985. Their analysis produces estimates of β on the order of 10%, which is much larger than the 2% typically found.

Endogeneity raises a second identification issue with respect to the relationship between β-convergence and economic convergence: this idea appears in Cohen (1996) and Goetz and Hu (1996). Focusing on the Solow regressors, the value of β can fail to illustrate how initial conditions affect expected future income differences if the population and saving rates are themselves functions of income. Hence, $\beta \geq 0$ may be compatible with at least partial economic convergence, if the physical and human capital savings rates depend, for example, on the level of income. In contrast, $\beta < 0$ may be compatible with economic divergence if the physical and human capital accumulation rates for rich and poor are diverging across time. As such, this critique is probably best understood as a debate over what variables are the relevant initial conditions for evaluating (22) and/or (23). Cohen (1996) argues that the conventional human capital accumulation equation, in which accumulation is proportional to per capita output, is misspecified, failing to account for feedbacks from the stock of human capital to the accumulation process. This feedback means that poor countries with low initial stocks of human capital fail to accumulate human capital as quickly as richer ones. Goetz and Hu (1996) directly focus on the feedback from income to human capital accumulation.

The implications of this form of endogeneity for empirical work on convergence are mixed. Cohen (1996) concludes that a proper accounting for the dependence of human capital accumulation on initial capital stocks reconciles conditional β – convergence with unconditional β – divergence for a broad cross-section. Goetz and Hu (1996), in contrast, find that estimates of the speed of convergence are increased if one accounts for the effect of income on human capital accumulation for counties in the US South. This seems to be an area that warrants much more work.

d. measurement error

As Abramovitz (1986), Baumol (1986), DeLong (1988), Romer (1990), and Temple (1998) point out, measurement errors will tend to bias regression tests towards results consistent with the hypothesis of β-convergence. This occurs because, by construction, $\gamma_{i,t}$ is measured with positive (negative) error when $\log y_{i,0}$ is measured with negative (positive) error so there tends to be a negative correlation between the measured values of the two variables even if there is none between the true values. To see this, we ignore the issue of control variables and consider the case where growth is described by $\gamma_i = k + \beta \log y_{i,0} + \varepsilon_i$ where ε_i is independent across observations. Suppose that log output is measured with error so that the researcher only observes $\varsigma_{i,t} = \log y_{i,t} + e_{i,t}$, $t = 0, T$ where $e_{i,t}$ is a serially uncorrelated random variable with variance σ_e^2 and distributed independently of $\log y_{i,s}$ and ε_i for all *i* and *s*. The regression of observed growth rates will, under these assumptions, obey the equation

$$T^{-1}\left(\varsigma_{i,T} - \varsigma_{i,0}\right) = k + \beta\varsigma_{i,0} + T^{-1}e_{i,T} - \left(\frac{\beta T + 1}{T}\right)e_{i,0} + \varepsilon_i \qquad (31)$$

This is a classic errors in variables problem; the term $\left(\frac{\beta T + 1}{T}\right)e_{i,0}$ is negatively correlated with $\varsigma_{i,0}$ which induces a negative bias in the estimate $\hat{\beta}$. In other words, the regression of observed growth rates on observed initial incomes will tend to produce an estimated coefficient that is consistent with the β-convergence hypothesis even if the hypothesis is not reflected in the actual behavior of growth rates across countries. In practice, as Temple (1998) explains, the direction of the bias is made ambiguous by the possibilities that the $e_{i,t}$ are serially dependent and that other right-hand-side (conditioning) variables are also measured with error. The actual effect of measurement error on results then becomes an empirical matter to be investigated by individual researchers.

In studying the role of the level of human capital in determining the rate of growth, Romer (1990) estimates a growth equation that has among its explanatory variables the level of per capita income at the beginning of the sample period. Consistent with the conditional β-convergence hypothesis, he finds a negative and significant coefficient on this variable when the equation is estimated by ordinary least squares. Wary of the possibility and effects of measurement error in initial income, as well as in the human capital variable – the literacy rate – Romer also estimates the equation using the number of radios per 1000 inhabitants and (the log of) per capita newsprint consumption as instruments for initial income and literacy with the result that the coefficients on both variables become insignificant "suggesting" that the OLS results are "attributable to measurement error" (p. 278).

Temple (1998) uses the measurement error diagnostics developed by Klepper and Leamer (1984), Klepper (1988), and classical method-of-moments adjustments, to investigate the effects of measurement error on the estimated rate of convergence in MRW's augmented Solow model. He finds that allowing for the possibility of small amounts of unreliability in the measurement of initial income implies a lower bound on the estimated convergence rate just above zero – too low to elevate conditional convergence to the status of a stylized fact. Barro and Sala-i-Martin (2004, pp. 472-3) use lagged values of state personal income as instruments for initial income to check for the possible effects of measurement error in their β-convergence tests for the US states. They find little change in the estimated convergence rates and conclude that measurement error is not an important determinant of their results. Barro (1991) follows the same procedure for other data sets and reaches a similar conclusion about the unimportance of measurement error in his results.

Some authors have attempted to address the sources of measurement error. Dowrick and Quiggin (1997) is a notable example in this regard in their consideration of the role of price indices in affecting convergence tests. Specifically, they examine the effect of constant price estimates of GDP on β-convergence calculations and find that when the prices used to construct these measures are based on prices in advanced economies, tendencies towards convergence are understated.

e. effects of linear approximation

There is a body of research that explores the effects of the approximations that are employed to produce the linear regression models used to evaluate β-convergence. As outlined earlier, regression tests of the β-convergence hypothesis rely on a log-linear approximation to the law of motion in a one sector neoclassical growth model. In addition to the possibility that Taylor series approximations in the nonstochastic version of the model are inadequate, Binder and Pesaran (1999) show that the standard practice of adding a random term to the log-linearized solution of a nonstochastic growth model does not necessarily produce the same behavior as associated with the explicit solution of a stochastic model.

Efforts to explore the limits of the linear approximation used in empirical growth studies have generally concluded that the approximation is reasonably accurate. Romer (2001, p. 25 n. 18) claims that the approximation will be "quite reliable" in this context and Dowrick (2004) presents results showing that the approximation to the true transition dynamics is quite good in a Solow model with a single capital good and an elasticity of output with respect to capital of $2/3$. This is larger than the typical physical capital share but it is not an unreasonable number for the sum of the shares of physical and human capital. To test for nonlinearity, Barro (1991) adds the square of initial (1960) income to one of his regressions and finds a positive estimated coefficient implying that the rate of convergence declines as income rises and that it is positive only for incomes below $10800 – a figure that exceeds all of the 1960 income levels in his sample. However, the t-ratio for the estimated coefficient on the square of initial income is just 1.4 which represents weak evidence against the adequacy of the approximation.

How should one interpret such findings? At one level, these studies conclude that the approximation used to derive the equation used in cross-section convergence studies appears to be reasonably accurate. It follows that the previously discussed nonlinearities in the growth process found by researchers investigating the possibility of multiple steady states do not reflect the inadequacy of the linear approximation used in most cross-section studies. Put differently, evidence of nonlinearity appears to reflect deeper factors than simple approximation error from the use of a first order Taylor series expansion.

iii. Distributional approaches to convergence

A second approach to convergence focuses on the behavior of the cross-section distribution of income in levels. Unlike the β-convergnece approach, the focus of this literature has been less on the question of relative locations within the income distribution, i.e. whether one can expect currently poor countries to either equal or exceed currently affluent countries, but rather on the shape of the distribution as a whole. Questions of this type naturally arise in microeconomic analyses of income inequality, in which one may be concerned with whether the gap between rich and poor is diminishing, regardless of whether the relative positions of individuals are fixed over time.

a. σ-convergence

Much of the empirical literature on the cross-country income distribution has focused on the question of the evolution of the cross-section variance of $\log y_{i,t}$. For a set of income levels let $\sigma^2_{\log y,t}$ denote the variance across i of $\log y_{i,t}$. σ-convergence is said to hold between times t and $t+T$ if

$$\sigma^2_{\log y,t} - \sigma^2_{\log y,t+T} > 0 \tag{32}$$

This definition is designed, like β-convergence, to formalize the idea that contemporary income differences are transitory, but does so by asking whether the dispersion of these differences will decline across time.

Recent work has attempted to identify regression specifications from which one can infer σ-convergence. Friedman (1992) and Cannon and Duck (2000) argue that it is possible to produce evidence concerning σ-convergence from regressions of the form

$$\gamma_i = T^{-1}\left(\log y_{i,t+T} - \log y_{i,t}\right) = \alpha + \pi \log y_{i,t+T} + \varepsilon_i \tag{33}$$

To see why this is so, following Cannon and Duck (2000), observe that $\sigma-$ convergence requires that $\sigma_{\log y_{i,t},\log y_{i,t+T}} < \sigma^2_{\log y_{i,t}}$. The regression coefficient in (33) may be written as

$$\pi = T^{-1}\left(1-\frac{\sigma_{\log y_{i,t},\log y_{i,t+T}}}{\sigma^2_{\log y_{i,t+T}}}\right) \tag{34}$$

which means that $\pi < 0$ implies $\sigma_{\log y_{i,t},\log y_{i,t+T}} < \sigma^2_{\log y_{i,t}}$. Postiveness definiteness of the variance/ covariance matrix for $\log y_{i,t}$ and $\log y_{i,t+T}$ requires that $\left(\sigma_{\log y_{i,t},\log y_{i,t+T}}\right)^2 < \sigma^2_{\log y_{i,t}}\sigma^2_{\log y_{i,t}}$ Therefore, if $\pi < 0$, then it must be the case that (32) holds. Hence a test that accepts null hypothesis that $\pi < 0$ by implication accepts the null hypothesis of σ-convergence. But even this type of test has some difficulties. As pointed out by Bliss (1999,2000), it is difficult to interpret tests of σ-convergence since these tests presume that the data generating process is not invariant; an evolving distribution for the data makes it difficult to think about test distributions under a null. Additional issues arise when unit roots are present.

One limitation to this approach is that it is not clear how one can formulate a sensible notion of conditional σ-convergence. A particular problem in this regard is that one would not want to control for initial income in forming residuals, which would render the concept uninteresting as it could be generated by nothing more than time-dependent heteroskedasticity in the residuals. On the other hand, omitting income would render the interpretation of the projection residuals problematic since initial income is almost certain to be correlated with the variables that have been included when the residuals are formed. An economically interesting formulation of conditional σ-convergence would be a useful contribution.

b. evolution of the world income distribution

Work on σ-convergence has helped stimulate the more general study of the evolution of the world income distribution. This work involves examining the cross-

section distribution of country incomes at two or more points in time in order to identify how this cross-section distribution has changed. Of particular interest in such studies is the presence or emergence of multiple modes in the distribution. Bianchi (1997) uses nonparametric methods to estimate the shape of the cross-country income distribution and to test for multiple modes in the estimated density. He finds evidence of two modes in densities estimated for 1970, 1980, and 1989. Moreover, he finds a tendency for the modes to become more pronounced and to move further apart over time. This evidence supports the ideas of a vanishing middle as the distribution becomes increasingly polarized into "rich" and "poor" and of a growing disparity between those two groups. While such polarization might be desirable, were it the case that middle income economies were becoming high income ones, Bianchi's evidence suggests that much of this movement represents a transition from middle income to poor. Further, by "cutting" each of the estimated densities at the anti-mode between the two modes, Bianchi is able to measure mobility within the distribution by counting the crossings of the cut points. These crossings represent countries moving from one basin of attraction to the other. Just 3 of the possible 238 crossings are observed.[29] The implication is that there is very little mobility within the cross-country income distribution. The 20 or so countries in the "rich" basin of attraction in 1970 are still there in 1989 and similarly for the 100 or so countries starting in the "poor" basin.

Paap and van Dijk (1998) model the cross-country distribution of per capita income as the mixture of a Weibull and a truncated normal density. The Weibull portion captures the left-hand mode and right skewness in the data while the truncated normal portion captures the right-hand mode. This combination is selected after testing the goodness of fit of various combinations of the normal density (truncated at zero), gamma, log normal and Weibull distributions; the data set that is employed measures levels of real GDP per capita for 120 countries for the time period 1960 and 1989. They find a bimodal fitted density in each year with "poor" and "rich" components corresponding to the Weibull and truncated normal densities respectively. The computed means of these

[29]Bianchi's data contains 119 countries observed at 3 distinct years, so each country is capable of making two crossings. The only crossings observed are 1) Trinidad and Tobago, which moves down between 1980 and 1989, 2) Venezuela, which moves down between 1970 and 1980, and 3) Hong Kong, which moves up between 1970 and 1980.

components diverge over the sample period and the weight given to the poor component in the mixture jumps in the mid-1970's from about .72 to about .82 implying that the mean gap between rich and poor countries grew and the poor increased in number. The attention to levels rather than log levels makes it hard to evaluate the welfare significance of this increased dispersion.

Recently, analyses of the distributions of income and growth have focused on identifying differences in these distributions across time and across subsets of countries. Anderson (2003) studies changes in the world income distribution by using nonparametric density function estimates combined with stochastic dominance arguments to compare the distributions at different points in time.[30] These methods allow him to construct measures of polarization of the income distribution; polarization is essentially characterized by shifts in probability density mass that increase disparities between relatively rich and relatively poor economies. Anderson finds that between 1970 and 1995 polarization between rich and poor countries increased throughout the time period. Maasoumi, Racine, and Stengos (2003) analyze the evolution of the cross-country distributions of realized, predicted, and residual growth rates; fitted growth rates and residuals are formed from nonparametric growth regressions using the Solow variables. These authors find that the distributions of growth rates for OECD and non-OECD countries are persistently different between 1965 and 1995, with the OECD distribution's variance reducing over time whereas the non-OECD distribution appears to be becoming less concentrated. One finds the same results for fitted growth rates; in contrast it is difficult to identify dimensions along which the distributions of OECD and non-OECD growth rate residuals differ. The major methodological difference between these papers relative to Paap and van Dijk (1998) is that these analyses do not rely on a mixture specification.

Distributional approaches suggest the utility of convergence measures that are based on the complete properties of probability measures characterizing output for different economies. Letting $\mu_i(x)$ and $\mu_j(x)$ denote the probability density functions

[30]Anderson (2004) discusses issues related to the interpretation and econometric implementation of these methods.

for the variable of interest in economies i and j respectively, Anderson and Ge (2004) propose computing the convergence statistic $CI_{i,j}$

$$CI_{i,j} = \int_{-\infty}^{\infty} \min\left(\mu_i(x), \mu_j(x)\right) dx \tag{35}$$

This statistic is bounded between 0 and 1; a value of zero means that the density functions never assign positive probability to any common intervals or values of x whereas a value of 1 means that the densities coincide on all positive probability intervals or values. Anderson and Ge (2004) refer to the case $CI_{i,j} = 1$ as complete convergence. This statistic differs from the convergence measure described by eq. (21) as it evaluates differences between current densities and not asymptotic ones, but they are clearly closely related.

In our view, this approach will likely prove useful in a range of contexts. In particular, if one is interested in comparing income distributions between two economies, the Anderson-Ge statistic is a natural metric. In growth contexts, it is less clear whether the higher moments that distinguish (22) from (35) are of major concern, at least in the context of current debates.

d. distribution dynamics

In a series of papers, Quah (1993a,b,1996a,b,c,1997) has persuasively criticized standard regression approaches to studying convergence issues for being unable to shed light on important issues of mobility, stratification, and polarization in the world income distribution. Rather than studying the average behavior of a representative country, Quah proposes a schema, which he calls "distribution dynamics", for studying the evolution of the entire cross-country income distribution. One way of implementing this approach is to assume that the process describing the evolution of the distribution is time-invariant and first-order Markov. Discretizing the state space then permits representation of cross-country income distribution as a probability mass function, λ_t, with an associated transition matrix, M. Each row of M is a probability mass function describing the

distribution over states of the system after one transition given that the system is currently in the state corresponding to that row. The evolution of the income distribution can then be described by $\lambda_t = M'\lambda_{t-1}$ so that $\lambda_{t+s} = (M^s)'\lambda_t$ is the *s*-step-ahead probability mass function and $\lambda_\infty = M'\lambda_\infty$ defines the long-run (ergodic) mass function (if it exists). Quah (1993b, 1996b) takes this approach and finds that the estimated *M* implies a bimodal ("twin-peaked") ergodic mass function indicating a tendency towards polarization in the evolution of the world income distribution.[31]

Updating Quah's analysis using more recent data, Kremer, Onatski, and Stock (2001) also find evidence of twin-peaks in the long-run distribution of per capita incomes. However, they find the rich (right-hand) peak to be much larger than the poor (left-hand) peak unlike Quah, who found similarly sized peaks at both ends of the distribution. Kremer, Onatski, and Stock's point estimates imply that most countries will ultimately move to the rich state although, during the transition period, which could last hundreds of years, polarization in the income distribution may worsen. They are also unable to reject the hypothesis that there is a single right-hand peak in the long-run distribution. Quah (2001) responds to these claims by arguing that the imprecision in the estimates of the ergodic distributions is such that it is not possible to reject a wide range of null hypotheses including, by construction, that of twin-peakedness. Importantly, as Quah notes his work and that of others, including Kremer, Onatski, and Stock, is consistent with the view that the global poor are many in number and likely to be so for a very long time.

In addition, as Quah (1996c,1997,2001) and Bulli (2001) discuss, the process of discretizing the state space of a continuous variable is necessarily arbitrary and can alter the probabilistic properties of the data. Especially relevant here is the fact that the shape of the ergodic distribution can be altered by changing the discretization scheme. Reichlin (1999) demonstrates that the dynamic behavior inferred from the analysis of Markov transition probabilities, and the apparent long-run implications of that behavior, are

[31]As Quah (1993b, footnote 4) explains, the estimated ergodic distributions "… should *not* be read as forecasts of what will happen in the future…" (his emphasis). Rather, he continues, they "… should be interpreted simply as characterizations of tendencies in the post-War history that actually realized."

sensitive to the discretization scheme employed; this work also shows that the estimated ergodic distribution can be sensitive to small changes in the transition probabilities. Bulli (2001) addresses this critique and shows how to discretize the state space in a way that preserves the probabilistic properties of the data. Applying her method to cross-country income data she finds an estimated ergodic distribution quite different from that found by arbitrary discretization as well as being an accurate approximation to the distribution computed using a continuous state space method.

An alternative formulation of distribution dynamics that avoids discretization problems is proposed by Quah (1996c,1997) and models the cross-country income distribution at time *t* with the density function, $f_t(x)$. If the process describing the evolution of the distribution is again assumed to be time-invariant and first-order Markov, then density at time $t+\tau$, $\tau > 0$, will be $f_{t+\tau}(x) = \int_0^\infty g_\tau(x|z) f_t(z)\, dz$ where $g_\tau(x|z)$ is the τ-period-ahead density of *x* conditional on *z*. The function $g_\tau(x|z)$ is the continuous analog of the transition matrix *M* and, assuming it exists, the ergodic (long-run) density function, $f_\infty(x)$, implied by $g_\tau(x|z)$ is the solution to $f_\infty(x) = \int_0^\infty g_\tau(x|z) f_\infty(z)\, dz$. Using nonparametric methods, Quah (1996c,1997) estimates various $g_\tau(x|z)$ and finds strong evidence of twin-peakedness in the cross-country income distribution. The estimated ergodic densities presented by Bulli (2001) and Johnson (2004) support Quah's conclusions.

Azariadis and Stachurski (2003) derive the form of the $g_\tau(x|z)$ implied by a stochastic version of the model in Azariadis and Drazen (1990). Estimation of the model's parameters enables them to compute forward projections of the sequence of cross-country income distributions, and ultimately the ergodic distribution, implied by the model. Consistent with the work of Quah (1996c, 1997) they find bimodality to be a pervasive feature of the sequence of distributions for about 100 years. Eventually, however, all countries transition to the rich mode so the ergodic distribution is unimodal as found by Kremer, Onatski, and Stock (2001). As Quah (2001) notes, there is "as yet" no theory of inference for this case so reconciliation of this result with the view that the

ergodic distribution is bimodal cannot be done through formal statistical tests. However, while Quah (2001) observes that such a theory is an "obvious next step," he suggests that we may be close to the limits of what can be reasonably inferred from the cross-country income data.

Johnson (2000) offers an interpretation of $g_\tau(x|z)$ which draws an analogy between the median of the conditional distribution and the law of motion of a non-stochastic one-variable dynamic system. The median is the function $m(x)$ such that $\int_0^{m(x)} g_\tau(z|x)dz = .5$ so that a country with income of $m(x)$ at time *t* has an equal chance of having a higher or lower income at time $t+\tau$. Consider a point x_0 such that $m(x_0) = x_0$ and suppose that, in some neighborhood of x_0, $m(x) > x$ for $x < x_0$ and $m(x) < x$ for $x > x_0$ implying $\Pr(x_{t+\tau} > x_t) > .5$ for $x < x_0$ and $\Pr(x_{t+\tau} < x_t) > .5$ for $x > x_0$ so that, in this neighborhood, countries with incomes different from x_0 tend to move toward x_0. In the long run we may expect to find many countries in the vicinity of x_0 creating the tendency for a mode in the ergodic density, $f_\infty(x)$, at x_0. Similarly, in a non-stochastic one-variable dynamic system with the law of motion $x_{t+\tau} = m(x_t)$, the condition on the phase diagram for the local stability of a steady-state at x_0 is that the graph of $m(x)$ intersects the 45° line from above at x_0. In both cases, x_0 is a point of accumulation in the sense that the long-run probability of finding countries in the vicinity of x_0 will tend to be high relative to that elsewhere. Conversely, just as steady states are unstable in the non-stochastic case when $m(x)$ crosses the 45° line from below, analogous points in the stochastic case tend to produce antimodes in the ergodic density.

While Quah's estimated $g_\tau(x|z)$ indicate a strong tendency towards polarization in the world income distribution, they do not reveal much about intra-distribution mobility. Bimodality is arguably of less concern in a normative sense if there is movement between the two modes than it is if there is none. Quah (1996c) studies the mobility within the distribution by computing, (through stochastic simulation) the mean

time for a "growth miracle" which he defines as passage from the 10th to 90th percentile of the distribution. He finds an expected time of 201 years for such a miracle to occur.

Quah's methods have subsequently been applied to a range of contexts. Andres and Lamo (1995) apply these methods to the OECD, Lamo (2000) to the regions of Spain, Johnson (2000) to US states, Bandyopadhyay (2002) to the Indian states, and Andrade, Laurini, Madalozzo, and Valls Pereira (2004) to Brazilian municipalities. These methods have also been extended to broader notions of distributional dynamics. Fiaschi and Lavezzi (2004) develop an analysis of the joint distribution of income levels and growth rates; their findings are compatible with the existence of multiple equilibria in the sense that countries may become trapped in the lower part of the income distribution.

e. relationship between distributional convergence and the persistence of initial conditions

Distributional methods have proven important in establishing stylized facts concerning the world income distribution. At the same time, there has been relatively little formal effort to explore the implications of findings such as twin peaks for the empirical salience of alternative growth theories. Some potential implications of distributional dynamics for evaluating theories are suggested by Quah (1996c), who finds that conditioning on measures of physical and human capital accumulation similar to those used by Mankiw, Romer, and Weil (1992) and a dummy variable for the African continent has little effect on the dynamics of the cross-country income distribution. The polarization and immobility features are similar in both cases and conditioning increases the expected time for a growth miracle to 760 years.[32] These results suggest that the heterogeneity revealed by the distributional approaches is, at least in part, due to the existence of convergence clubs.

[32]Other efforts to find determinants of polarization and immobility have produced mixed results. For the OECD countries, Andres and Lamo (1995) condition on the steady state implied by the Solow model and find little decrease in the tendency to polarization unless country specific effects are permitted. Lamo (2000) finds only a small increase in mobility for Spanish regions after conditioning on interregional migration flows. Bandyopadhyay (2002) shows that infrastructure spending and education measures appear to contribute to polarization between rich and poor states of India.

That being said, in general, it is relatively difficult to interpret properties of the cross-country income distribution in the context of economic convergence in the sense of (22). To see why this is so, it is useful to focus on the absence of a clear relationship between β-convergence, which measures the relative growth of rich versus poor countries and σ-convegence, which focuses explicitly on the distribution of countries. These two convergence notions do not have any necessary implications for one another, i.e. one may hold when the other does not. For our purposes, what is important is that σ-convergence is not an implication of β-convergence and so does not speak directly to the question of the transience of contemporary income differences. The erroneous assertion that β-convergence implies σ-convergence is known as Galton's fallacy and was introduced into the modern economic growth context by Friedman (1992) and Quah (1993a).

To understand the fallacy, suppose that log per capita output in each of *N* countries obeys the AR(1) process

$$\log y_{i,t} = \alpha + \varsigma \log y_{i,t-1} + \varepsilon_{i,t} \tag{36}$$

where $0 < \varsigma < 1$ and the random variables $\varepsilon_{i,t}$ are i.i.d across countries and time. For this model, each country will, by definition (22), exhibit convergence as any contemporaneous difference in output between two countries will disappear over time. Further, it is easy to see, using $\gamma_i = T^{-1}\left(\log y_{i,t+T} - \log y_{i,t}\right)$, that the regression of growth on a constant and initial income will exhibit β-convergence. This is immediate when one considers growth between t and $t+1$ which means that growth obeys

$$\gamma_{i,t} = \alpha + \left(\varsigma - 1\right)\log y_{i,t-1} + \varepsilon_{i,t} \tag{37}$$

where $\varsigma - 1 < 0$ by assumption. In this model, by construction, the unconditional population variance of log output is constant because the reduction in cross-section variance associated with the tendency of high-income countries to grow more slowly than low-income countries is offset by the presence of the random shocks $\varepsilon_{i,t}$. This indicates

why σ-convergence is not a natural implication of long run independence from initial conditions; rather σ-convergence captures the evolution of the cross-section income distribution towards an invariant measure. This suggests that an important next step in the distributional approach to convergence is the development of tools which will allow distribution methods to more directly adjudicate substantive growth questions as they relate to the persistence of initial conditions.

iv. time series approaches to convergence

A final approach to convergence is based on time series methods. This approach is largely statistical in nature, which allows various hypotheses about convergence to be precisely defined, and thereby reveals appropriate strategies for formal testing. A disadvantage of the approach is that it is not explicitly tied to particular growth theories. Bernard and Durlauf (1995,1996), Evans (1998) and Hobijn and Franses (2000) provide a systematic framework for time series convergence tests.

Following Bernard and Durlauf (1995), a set of countries I is said to exhibit convergence if

$$\lim_{T\to\infty} \text{Proj}\left(\log y_{i,t+T} - \log y_{j,t+T} \,|F_t\right) = 0 \;\forall i, j \in I \tag{38}$$

where $\text{Proj}\left(a|b\right)$ denotes the projection of a on b and F_t denotes some information set; operationally, this information set will typically contain various functions of time and current and lagged values of $\log y_{i,t}$ and $\log y_{j,t}$. Relative to our previous discussion, this definition represents a form of unconditional convergence that is closely related to (22). One can modify the definition to apply to the residual of per capita income after it has been projected on control variables such as savings rates in order to produce a definition of conditional convergence, but this has apparently not been done in the empirical literature.

In evaluating (38), researchers have generally focused on whether deterministic or stochastic trends are present in $\log y_{i,t} - \log y_{j,t}$; the presence of such trends immediately

implies a violation of (38). As such, time series tests of convergence have typically been implemented using unit root tests. One reason for this focus is that the presence of unit roots in $\log y_{i,t} - \log y_{j,t}$ allows for an extreme and therefore particularly interesting form of divergence between economies since a unit root implies that the difference $\log y_{i,t} - \log y_{j,t}$ will, with probability 1, become arbitrarily large at some point in the future.

The use of unit root and related time series tests has important implications for the sorts of countries that may be tested. Time series tests presuppose that $y_{i,t}$ may be thought of as generated by an invariant process in either levels or first differences, i.e., either levels or first differences may be modeled as the sum of deterministic terms plus a Wold representation for innovations. Such an assumption has significant economic content. As argued by Bernard and Durlauf (1996) countries that start far from their invariant distributions and are converging towards them, as occurs for countries that are in transition to the steady-state in the Solow-Swan model, will be associated with $\log y_{i,t} - \log y_{j,t}$ series that do not fulfill this requirement. Hence, tests of (38) can produce erroneous results if applied to such economies. To see this intuitively, suppose that for country *i*, $\log y_{i,t} = \log y_{i,t+1}$ for all *t*, so that country has converged to a constant steady-state. Suppose that country *j* has the same steady-state as country *i* and is monotonically converging to this state so that $\log y_{i,t} > \log y_{j,t}$ for all observations. Then $\log y_{i,t} - \log y_{j,t} > 0$ for all *t* in the sample; which means that the series has a nonzero mean and tests that fail to account for the fact that the density of $\log y_{i,t} - \log y_{j,t}$ is changing across time can easily give erroneous inferences. For example one may use a test and conclude $\log y_{i,t} - \log y_{j,t}$ possesses a nonzero mean and erroneously interpret this as evidence against convergence, when the fact that the process does not have a time-invariant mean is ignored. This argument suggests that time series convergence tests are really only appropriate for advanced economies that may plausibly be thought of as characterized by invariant distributions.

Generally, the first generation of these tests rejected convergence for countries as well as other economic units. For example, Bernard and Durlauf (1995), studying 15

advanced industrialized economies between 1900 and 1989 based on data developed in Maddison (1982,1989), find little evidence that convergence is occurring; Hobijn and Franses (2000) similarly find little evidence of convergence across 112 countries taken from the Penn World Table for the period 1960-1989. The findings of nonconvergence in output levels are echoed in recent work by Pesaran (2004a) who employs convergence definitions that explicitly focus on the probability of large deviations, i.e. eq. (25). He finds little evidence of output level convergence using either the Maddison or Penn World Table data.

Relatively little explicit attention has been paid to the question of systematically identifying convergence clubs using time series methods. One exception is Hobijn and Franses (2000) who employ a clustering algorithm to identify groups of converging countries.[33] Their algorithm finds many small clusters in their sample of 112 countries – depending on the particular rule used to determine cluster membership, they find 42 or 63 clusters with most containing just two or three countries. Hobijin and Frances view these clusters as convergence clubs but it is not clear that they represent groups of countries in distinct basins of attraction of the growth process. Absent controls for structural characteristics, these groupings could simply reflect the pattern of differences in those characteristics rather than differences in long-run outcomes due to differences in initial conditions. Moreover, the Bernard and Durlauf (1996) argument about the substantive economic assumptions that underlie time series methods for studying convergence seems applicable here. Given the breadth of the sample used by Hobijn and Franses, it is unlikely that it contains only data generated by countries whose behavior is near their respective steady-states; such an assumption is much more plausible for restricted samples such as the OECD countries. The clusters they find could thus reflect, in many cases at least, transition dynamics rather than convergence clubs. An important extension of this work would be the exploration of how one can distinguish convergence clubs from

[33]Corrado, Martin, and Weeks (2004) extend this approach to allow for time variation in clusters. They conclude that there is substantial evidence of club convergence as opposed to overall convergence for European regions. A nice feature of their analysis is the effort to interpret the clubs that are identified statistically with alternative economic theories, and conclude that geographic proximity and demographic similarity correlate with their observed clusters.

what may be called "transition" clubs, i.e. groups of countries exhibiting similar transition dynamics.

A number of studies of time series convergence have criticized these claims of nonconvergence; these criticisms are based upon inferential issues that have arisen in the general unit roots literature. One of these issues concerns the validity of unit root tests in the presence of structural breaks in $\log y_{i,t} - \log y_{j,t}$; as argued initially by Perron (1989), the failure to allow for structural breaks when testing for unit roots can lead to spurious evidence in support of the null hypothesis that a unit root is present. An initial analysis of this type in cross-country contexts is Greasley and Oxley (1997) who, imposing breaks exogenously, find convergence for Denmark and Sweden whereas the sort of test employed by Bernard and Durlauf (1995) does not. The role of breaks in time series convergence tests is systematically studied in Li and Papell (1999). An important feature of their analysis is that Li and Papell avoid exogenous imposition of trend breaks and in fact find that the dates of these breaks exhibit some heterogeneity, although many of them cluster around World War II. Li and Papell find that the evidence for OECD convergence is more mixed than did Bernard and Durlauf (1995) in the sense that allowing for trend breaks reduces the number of country pairs that fail to exhibit convergence. Related findings are due to Carlino and Mills (1993) who study US regions and reject convergence except under specifications that allow for a trend break in 1946. These conclusions are shown by Loewy and Papell (1996) to hold even if one allows potential trend breaks to be endogenously determined by the data.

While the analysis of trend breaks and convergence tests is valuable because of its implications about the time series structure of output differences between countries, studies of this type suffer from some interpretation problems. The presence of the regime break is presumably suggestive of an absence of convergence in the sense of (22) or (38), since it implies that there is some component of $\log y_{i,t} - \log y_{j,t}$ that will not disappear over a sufficiently long time horizon. The time series definition of convergence is violated by any long-term predictability in output differences. Hence, claims by authors that allowing for data breaks produces evidence of convergence begs the question of what is meant by convergence. That being said, the sort of violation of (22) or (38) implied by a trend break is different from the type implied by a unit root. In particular, a break

associated with the level of output means that the output difference between two countries is always bounded, unlike the unit root case.

A distinct line of criticism of time series convergence tests is due to Michelacci and Zaffaroni (2000) who argue that convergence tests based on the presence of unit roots may perform badly when the true processes exhibit long memory. Let $\gamma(L)u_{i,j.t}$ denote the moving average representation for $\log y_{i,t} - \log y_{j,t}$. Suppose that the k'th coefficient in the representation has the property that

$$\gamma_k \propto k^{d-1},\ 0 < d < 1 \tag{39}$$

In this case, shocks die out at a hyperbolic rather than geometric rate, which is one definition of long memory in a time series process. Michelacci and Zaffaroni (2000) show that if output deviations exhibit long memory, one can reconcile the claim of β-convergence with time series evidence of divergence, i.e. the failure of various tests to reject the presence of a unit root in per capita output deviations. This is a potentially important reconciliation of these two distinct testing strategies.

That being said, the plausibility of a long memory characterization has yet to be established in the economics literature. One problem is that there is an absence of a body of economic theory that predicts the presence of long memory.[34] The existing theoretical justifications of long memory processes derive from aggregation arguments originating with Granger (1980); the conditions under which aggregation produces long memory do not have any particular empirical justification. In addition, there are questions concerning the ability of conventional statistical methods to allow one to distinguish between long memory models and various alternatives. Diebold and Inoue (2001) indicate how long memory may be spuriously inferred for series subject to regime shifts, so the strength of evidence of long memory cited by Michelacci and Zaffaroni (2000)

[34]There are at least two reasons why unit roots stem naturally from exisiting economic theories. First, technology shocks are generally modeled as permanent. Second, Euler equations often produce unit root or near unit root like conditions. The random walk theory of stock prices is one example of this, in which risk neutral agents produce unpredictability of stock price changes as an equilibrium.

may be questioned. Nevertheless, the Michelacci-Zaffaroni argument is important, not least because it focuses attention on the role in growth empirics of size and power issues that arise in all unit root contexts.

Time series approaches to convergence are melded with analysis related to σ-convergence in Evans (1996) who considers the cross-section variance of growth rates at time t,

$$\sigma_t^2 = \frac{1}{N}\sum_i \left(\log y_{i,t} - \overline{\log y}_t\right)^2 \tag{40}$$

where $\overline{\log y}_t = \frac{1}{N}\sum_i \log y_{i,t}$ and N is the cardinality of I. Evans observes that σ_t^2 may be represented as a unit root process with a quadratic time trend when there is no cointegration among the series $\log y_{i,t}$. This leads Evans to suggest a time-series test of convergence based on unit root tests applied to σ_t^2. Employing this test, Evans concludes that there is convergence to a common trend among 13 industrial countries. One interpretation problem with this analysis is that it allows different countries to possess different deterministic trends in per capita output albeit with the same trend growth rate. Such differences are obviously germane with respect to convergence as an economic concept being consistent, for example, with the club and conditional convergence hypotheses but not with the unconditional convergence hypothesis. Evans (1997) provides a time series approach to estimating rates of convergence. He shows that OLS applied to equation (18) yields a consistent estimator of β, and hence the rate of convergence, only if (i) each $\log y_{i,t} - \overline{\log y}_t$ obeys an AR(1) process having the same AR(1) parameter lying strictly between 0 and 1; and, (ii) the control variables, X_i and Z_i, account for all cross-country heterogeneity. He argues that neither condition is likely to hold and offers an alternative method of measuring the rate of convergence based on the supposition that $\log y_{i,t} - \overline{\log y}_t$ follows an AR(q) process with lag polynomial $\Lambda(L)$. Again, this specification allows countries to follow different parallel balanced

growth paths and Evans defines the rate of convergence for economy *i* as the rate at which $\log y_{i,t}$ "is expected to revert toward its balanced growth path far in the future." He shows that, given that it is a real, distinct, positive fraction, the dominant root of the polynomial $z^q \Lambda\left(z^{-1}\right)$ equals one minus this rate. Evans computes estimates of the convergence rates and their 90% confidence intervals for a sample of 48 countries over the period 1950-90 and for the contiguous US states over the period 1929-91. For the states, about a third of the point estimates are negative and about two-thirds of the confidence intervals contain zero, while for the countries, about half of the point estimates are negative and all but two of the confidence intervals contain zero. However, in spite of these positive estimated average convergence rates of 15.5% and 5.9% respectively, Evans' analysis fails to yield persuasive evidence in favor of the conditional convergence hypothesis since, in most cases, the hypothesis of a convergence rate of zero cannot be rejected at the 10% level of significance.

Later sections of the chapter will discuss how growth researchers can draw on time series data in other ways. One popular route has been to use panel data, with repeated observations on each country or region. Another method is to use techniques broadly similar to those of event studies in empirical finance, and trace out the consequences of specific events, such as major political or economic reforms. We will consider these approaches in Section VI.iii below.

v. sources of convergence or divergence

Abramowitz (1986), Baumol (1986), DeLong (1988) and many others, both before and since, view convergence as the process of follower countries "catching up" to leader countries by adopting their technologies. Some more recent contributors, such as Barro (1991) and Mankiw, Romer, and Weil (1992), adopt the view that convergence is driven by diminishing returns to factors of production.[35] In the neoclassical model, if

[35] When an economy is below its steady-state value of capital per efficiency unit of labor, the marginal product of capital is relatively high (and is higher than in the steady state). As a result, a given investment rate translates into relatively high output growth. Capital grows as well but, because of diminishing returns, the capital-output ratio rises and the marginal product of capital declines, causing the growth of output and capital to slow.

each country has access to the same aggregate production function the steady-state is independent of an economy's initial capital and labor stocks and hence initial income. In this model, long-run differences in output reflect differences in the determinants of accumulation, not differences in the technology used to combine inputs to produce output. Mankiw (1995, p. 301), for example, argues that for "understanding international experience, the best assumption may be that all countries have access to the same pool of knowledge, but differ by the degree to which they take advantage of this knowledge by investing in physical and human capital." Even if one relaxes the assumption that countries have access to the same production function, convergence in growth rates can still occur so long as each country's production function is concave in capital per efficiency unit of labor and each country experiences the same rate of labor-augmenting technical change.

Klenow and Rodríguez-Clare (1997a) challenge this "neoclassical revival" with results suggesting that differences in factor accumulation are, at best, no more important than differences in productivity in explaining the cross-country distribution of output per capita. They find that only about half of the cross-country variation in the 1985 level of output per worker is due to variation in human and physical capital inputs while a mere 10% or so of the variation in growth rates from 1960 to 1985 reflects differences in the growth of these inputs. The differences between the results of Mankiw, Romer, and Weil (1992) and the findings of Klenow and Rodríguez-Clare (1997a) in their reexamination of Mankiw, Romer and Weil have two principal origins. First, citing concerns about the endogeneity of the input quantities, Klenow and Rodríguez-Clare (1997a) eschew estimation of the capital shares and choose to impute parameters based on the results of other studies. Second, they modify Mankiw, Romer and Weil's measure of human capital accumulation by supplementing secondary school enrollment rates using data on primary enrollment. This yields a measure of human capital accumulation with less cross-country variation than that used by Mankiw, Romer and Weil. This one modification decreases

---

Eventually, the economy converges to a steady state in which capital and output grow at the same rate and the marginal product of capital is sustained at a constant level by labor-augmenting technical progress. Dowrick and Rogers (2002) find that both diminishing returns and technology transfer are important contributors to the convergence process. See also Bernard and Jones (1996) and Barro and Sala-i-Martin (1997).

the relative contribution of cross-country variation in human and physical capital inputs to variation in the 1985 level of output per worker to 40% from the 78% found by Mankiw, Romer, and Weil. Prescott (1998) and Hall and Jones (1999) confirm the view that differences in inputs are unable to explain observed differences in output and Easterly and Levine (2001, p. 177) state that "[t]he 'residual' (total factor productivity, TFP) rather than factor accumulation accounts for most of the income and growth differences across countries."

Unlike many authors, who estimate TFP as a residual after assuming a common Cobb-Douglas production function, Henderson and Russell (2004) use a non-parametric production frontier approach (data envelopment analysis) to decompose the 1965 to 1990 growth of labor productivity into (i) shifts in the (common, worldwide) production frontier (technological change); (ii) movements toward (or away from) the frontier (technological catch-up); and, (iii) capital accumulation. They find a dominant role for capital accumulation in the growth of the cross-country mean of labor productivity with human and physical capital each accounting for about half of that role.[36] They also observe that the distribution of labor productivity became more dispersed from 1965 to 1990 and their results suggest that physical and human capital accumulation were largely responsible for the increased dispersion.

The results of Henderson and Russell (2004) and those of the previous authors are, however, more consistent than it may seem. Klenow and Rodríguez-Clare (1997a), Hall and Jones (1999) and Barro and Sala-i-Martin (2004) argue that the standard growth accounting decomposition overstates the contribution of capital accumulation to output growth by attributing to capital the effect on output of increases in capital induced by increases in TFP. This effect also applies to Henderson and Russell's approach and adjusting for it provides some reconciliation of their findings with those of Klenow and Rodríguez-Clare (1997a), Prescott (1998) and Hall and Jones (1999). The standard growth accounting formula attributes a fraction (equal to labor's share of output) of the

[36]Note that any misspecification of the production function due to the Cobb-Douglas assumption in other studies will tend to increase the apparent variation in TFP relative to that found by Henderson and Russell (2004) under the weaker assumption of constant returns to scale. In a rare effort to evaluate the Cobb-Douglas specification, Duffy and Papageorgiou (2000) reject it in favor of a more general CES functional form.

growth in output per worker to growth in TFP and a fraction (equal to capital's share of output) to capital accumulation despite the fact that, in the steady-state, growth in output per worker is entirely due to technological progress (Barro and Sala-i-Martin (2004, p. 457-60) and Klenow and Rodríguez-Clare (1997a p. 75, fn. 4)). The total effect of technological progress on output growth can thus be estimated by dividing labor's share into the estimated growth rate of TFP. Interpreting "capital" broadly, labor's share is about $1/3$ suggesting that this effect is about three times the rate of growth of TFP. Henderson and Russell (2004, Table 5, row (a)) find that, on average, about 90% of the increase in output per worker over the 1965 to 1990 period is attributable to the accumulation of human and physical capital with increases in TFP accounting for the remaining 10%. Applying the adjustment discussed above suggests that technological progress accounts for about 30% of the growth in output per worker over this period while capital accumulation, due to transition dynamics, accounts for the remainder.

As well as determining the relative contributions of inputs and TFP to the cross-country variation in output and output growth, some have studied what features of the cross-country output distribution are explained by the cross-country distributions of inputs and TFP. Henderson and Russell (2004) document the emergence of a second mode in the cross-country distribution of output per worker between 1965 and 1990 and find changes in efficiency (the distance from the world technological frontier) to be largely responsible. A primary role for TFP in determining the shape of the long-run distribution of output per capita is found by Feyrer (2003) who uses Markov transition matrices estimated with data from 90 countries over the period 1970 to 1989 to estimate the ergodic distributions of output per capita, the capital-output ratio, human capital per worker, and TFP. He finds that the long-run distributions of both output per capita and TFP are bimodal while those of both the capital-output ratio and human capital per worker are unimodal. This result, Feyrer observes, has potentially important implications for theoretical modelling of development traps. It suggests that models of multiple equilibria that give rise to equilibrium differences in TFP are more promising than

models that emphasize indeterminacy in capital intensity or educational attainment.[37] It is also consistent with Quah's (1996c) finding that conditioning on measures of physical and human capital accumulation (and a dummy variable for the African continent) has little effect on the dynamics of the cross-country income distribution.

As discussed in Section III.iii.d, the shapes of ergodic distributions computed from transition matrices estimated with discretized data are not, in general, robust to changes in the way in which the state space is discretized. To avoid these problems, Johnson (2004) extends Feyrer's analysis using Quah's (1996c,1997) continuous state-space methods and finds evidence of bimodality in the long-run distributions of both the capital-output ratio and TFP in addition to that in the long-run distribution of output per capita. This finding is broadly consistent with data produced by a version of the Solow growth model that includes a threshold externality à la Azariadis and Drazen (1990) but may be partly due to the computation of TFP after supposing a Cobb-Douglas production function across countries. Accordingly, some care must be exercised when drawing conclusions from these results.

More generally, in much of the development accounting literature cited above, TFP is measured as a residual under the assumption of a concave worldwide production function. Durlauf and Johnson (1995) present evidence contrary to that assumption and in support of the implied multiple steady states in the growth process. It seems likely that the imposition of a concave production function in this case will tend to exaggerate the measured differences in TFP and so confound inferences about the importance of TFP variation.[38] While Henderson and Russell (2004)'s approach is nonparametric and free from any assumption of a particular technology *per se*, it estimates the world technology frontier by fitting a convex cone to data on outputs and inputs. The imposed convexity of the production set prevents the method from discovering any nonconvexities that may exist and, in addition to masking the presence of multiple steady states, convexifying

[37]Romer (1993) discusses the intellectual origins of the centrality of capital accumulation in models of economic development and argues that "idea gaps are central to the process of economic development" (p. 548).

[38]Graham and Temple (2003) show that the existence of multiple steady states can increase the variance and accentuate bimodality in the observed cross-country distribution of TFP.

these nonconvexities would tend to overstate the cross-country variation in TFP. The extent to which our current understanding of the relative contributions of variation in inputs and variation in TFP to the observed variation in income levels is influenced by the effects on measured TFP of a misspecified worldwide technology remains an open research question.

Despite these concerns and the differences in the precise estimates found by different researchers, it is clear that cross-country variation in inputs falls short of explaining the observed cross-country variation in output. The result that the TFP residual, a "measure of our ignorance" computed as the ratio of output to some index of inputs, is an important (perhaps the dominant) source of cross-country differences in long-run economic performance is useful but hardly satisfying and the need for a theory of TFP expressed by Prescott (1998) is well founded. Research such as Acemoglu and Zilibotti (2001) and Caselli and Coleman (2003) are promising contributions to that agenda.

V. Statistical models of the growth process

While the convergence hypothesis plays a uniquely prominent role in empirical growth studies, it by no means represents the bulk of empirical growth studies. The primary focus of empirical growth papers may be thought of as a general exploration of potential growth determinants. This work may be divided into three main categories: 1) studies designed to establish that a given variable does or does not help explain cross-country growth differences, 2) efforts to uncover heterogeneity in growth and 3) studies that attempt to uncover nonlinearities in the growth process. While analyses of these types are typically motivated by formal theories, operationally they represent efforts to develop statistical growth models that are consistent with certain types of specification tests.

Section V.i discusses the analysis of how specific determinants affect growth. We describe the range of different variables that have appeared in growth regressions and consider alternative methodologies for analyzing growth models in the presence of

uncertainty about which regressors should be included to define the "true" growth model. Section V.ii addresses issues of parameter heterogeneity. The complexity of the growth process and the plethora of new growth theories suggest that the mapping of a given variable to growth is likely a function of both observed and unobserved factors; for example, the effect of human capital investment on growth may depend on the strength of property rights. We explore methods to account for parameter heterogeneity and consider the evidence that has been adduced in support of its presence. Section V.iii focuses on the analysis of nonlinearities and multiple regimes in the growth process. Endogenous growth theories are often highly nonlinear and can produce multiple steady states in the growth process, both of which have important implications for econometric practice. This subsection explores alternative specifications that have been employed to allow for nonlinearity and multiple regimes and describes some of the main findings that have appeared to date.

i. specifying explanatory variables in growth regressions

In the search for a satisfactory statistical model of growth, the main area of effort has concerned the identification of appropriate variables to include in linear growth regressions, this generally amounts to the specification of Z in equation (18). Appendix 2 provides a survey of different regressors that have been proposed in the growth literature with associated studies that either represent the first use of the variable or a well known use of the variable.[39] The table contains 145 different regressors, the vast majority of which have been found to be statistically significant using conventional standards.[40] One reason why so many alternative growth variables have been identified is due to questions of measurement. For example, a claim that domestic freedom affects growth leaves unanswered how freedom is to be measured. We have therefore organized the body of growth regressors into 43 distinct growth "theories" (by which we mean conceptually

[39]Our choices of which studies to include should not be taken to reflect any stance on any cases where there is disagreement about priority as to who first proposed a variable.
[40]Of course, the high percentage of statistically significant growth variables reflects publication bias as well as data mining.

distinct growth determinants); each of these theories is found to be statistically significant in at least one study.

As Appendix 2 indicates, the number of growth regressors that have been identified approaches the number of countries available in even the broadest samples. And this regressor list does not consider cases where interactions between variables or nonlinear transformations of variables have been included as regressors; both of which are standard ways of introducing nonlinearities into a baseline growth regression. This plethora of potential regressors starkly illustrates one of the fundamental problems with empirical growth research, namely, the absence of any consensus on which growth determinants ought to be included in a growth model. In this section, we discuss efforts to address the question of variable choice in growth models.

To make this discussion concrete, define S_i as the set of regressors which a researcher always retains in a regression and let R_i denote additional controls in the regression, so that

$$\gamma_i = \psi S_i + \pi R_i + \varepsilon_i \tag{41}$$

Notice that the inclusion of a variable in *S* does not mean the researcher is certain that it influences growth, only that that it will be included in all models under consideration. To make this concrete, one can think of an exercise in which one wants to consider the relationship between initial income and growth. A researcher may choose to include initial income and the other Solow growth regressors in every specification of the model, but may in contrast be interested in the effects of different non-Solow growth regressors on inferences about the initial income/growth connection.

If one takes the regressors that comprise *R* as fixed, then statements about elements of ψ are straightforward. A frequentist approach to inference will compute an estimate of the parameter $\hat{\psi}$ with an associated distribution that depends on the data generating process; Bayesian approaches will compute a posterior probability density of ψ given the researcher's prior, the data, and the assumption that the linear model is correctly specified, i.e. the choice of variables in *R* corresponds to the "true" model. Designating

the available data as D and a particular model as m, this posterior may be written as $\mu(\psi|D,m)$.

The basic problem in developing statistical statements either about $\hat{\psi}$ or $\mu(\psi|D,m)$ is that there do not exist good theoretical reasons to specify a particular model m. This is *not* to say that the body of growth theories may not be used to identify candidates for *R*. Rather, the problem is that growth theories are, using a phrase due to Brock and Durlauf (2001a), openended. Theory openendedness means that the growth theories are typically compatible with one another. For example, a theory that institutions matter for economic growth is not logically inconsistent with a theory that emphasizes the role of geography in growth. Hence, if one has a set of K potential growth theories, all of which are logically compatible with one another (and all subsets of theories), there exist $2^K - 1$ potential theoretical specifications of the form (41), each one of which corresponds to a particular combination of theories.

One approach to resolving the problem of model uncertainty is based on identifying variables whose empirical importance is robust across different model specifications. This is the idea behind Levine and Renelt's (1992) use of extreme bounds analysis (Leamer (1983) and Leamer and Leonard (1983)) to assess growth determinants. To see how extreme bounds analysis may be applied to the assessment of robustness of growth determinants, suppose that one has specified a space of possible models M. For model m, the growth process is

$$\gamma_i = \psi_m S_i + \pi_m R_{i,m} + \varepsilon_{i,m} \tag{42}$$

where the subscripts *m* reflect the model specific nature of the parameters and associated residuals. One can compute $\hat{\psi}_m$ for every model in M. Motivated by Leamer (1983), Levine and Renelt employ the rule that there is strong evidence that a given regressor in *S*, call it s_l, robustly affects growth if the sign of the associated regression coefficient $\hat{\psi}_{l,m}$ is constant and the coefficient estimate is statistically significant across all model

specifications in M. In this analysis the S vector is composed of a variable of interest and other variables whose presence is held fixed across specifications.

In the Levine and Renelt (1992) analysis, S includes a constant, the initial income, the investment share of GDP, secondary school enrollment rates, and population growth; these variables proxy for those suggested by the Solow model. Models are distinguished by alternative combinations of 1 to 3 variables taken from a set of 7 variables; these correspond to alternative choices of $R_{i,m}$. Based on the constant sign and statistical significance criteria, Levine and Renelt (1992) conclude that the only robust growth determinants among the elements of S_i are initial inome and the share of investment in GDP. These two findings are confirmed in subsequent work by Kalaitzidakis, Mamuneas, and Stengos (2000) who allow for potential nonlinearities in (41). Specifically, they consider partially linear versions of (41), so that

$$\gamma_i = \psi_m S_i + f_m\left(\pi R_{i,m}\right) + \varepsilon_{i,m} \tag{43}$$

Note that the function $f(\cdot)$ is allowed to vary across specifications of *R*. As in Levine and Renelt (1992), Kalaitzidakis, Mamuneas, and Stengos conclude that initial income and physical capital investment rates are robust determinants of growth. Unlike Levine and Renelt, they also find that inflation volatility and exchange rate distortions are robust; this is interesting as it is an example where the failure to account for nonlinearity in one set of variables masks the importance of another.

From a decision-theoretic perspective, the extreme bounds approach is a problematic methodology. The basic difficulty, discussed in detail in Brock and Durlauf (2001a) and Brock, Durlauf, and West (2003) is that if one is interested in ψ_l because one is considering whether to change $s_{i,l}$, by one unit, i.e. one is advising country *i* on a policy change, the extreme bounds standard corresponds to a very risk averse way of responding to model uncertainty. Specifically, suppose that for a policymaker, $El\left(s_{i,l}, m\right)$ represents the expected loss associated with the current policy level in country *i*. We assume that one is only interested in the case where an increase in the policy raises

growth, which means we will assume that it is necessary for $\hat{\psi}_{l,m} > 0$ in order to conclude that one should make the change. One can approximate the *t*-statistic rule, i.e. requiring that the coefficient estimate for s_l be statistically significant in order to justify a policy as implying that

$$El\left(s_{i,l}+1,m\right)-El\left(s_{i,l},m\right)=\left(\hat{\psi}_{l,m}-2sd\left(\hat{\psi}_{l,m}\right)\right)>0 \tag{44}$$

where $sd\left(\hat{\psi}_{l,m}\right)$ is the estimate of the standard deviation associated with $\hat{\psi}_{l,m}$ and the statistical significance level required for the coefficient is assumed to correspond to a *t*-statistic of 2. This loss function may look odd, but it is in fact the sort of loss function implicitly assumed whenever one relies on *t*-statistics to make policy decisions. Extreme bounds analysis requires that (44) holds for every model in *M*. This requires that $El\left(s_{i,l}\right)$, the expected loss for a policymaker when one conditions only on the policy variable, has the property that

$$El\left(s_{i,l}+1\right)-El\left(s_{i,l}\right)>0 \Rightarrow El\left(s_{i,l}+1,m\right)-El\left(s_{i,l},m\right)>0 \ \forall m \tag{45}$$

This means that the policymaker must have minimax preferences with respect to model uncertainty, i.e. he will make the policy change only if it yields a positive expected payoff under the least favorable model in the model space. While there are reasons to believe that in practice, individuals assess model uncertainty differently than within-model uncertainty[41], the extreme risk aversion embedded in (45) seems hard to justify.

Even when one moves away from decision-theoretic considerations, extreme bounds analysis is somewhat difficult to interpret as a statistical procedure. Hoover and Perez (2004), for example, show that the use of extreme bounds analysis can lead to the conclusion that many growth determinants are fragile even when they are part of the data

[41]See discussion in Brock, Durlauf, and West (2003) of the Ellsberg Paradox.

generating process. They also find that the procedure has poor power properties in the sense that some regressors that do not matter may spuriously appear to be robust.[42]

The concern that extreme bounds analysis represents an excessively conservative approach to evaluating empirical results led Sala-i-Martin (1997a,b) to propose a different way to evaluate the robustness of findings. Within a model, suppose there is an evaluative criterion for $\hat{\psi}_m$ that is used to determine whether the variable s_l matters for the growth process. One example of such a standard is whether or not $\hat{\psi}_{l,m}$ is statistically significant at some level. Sala-i-Martin first proposes averaging the statistical significance levels via

$$\hat{S}_i = \sum_m \hat{\omega}_m \hat{S}_{i,m} \tag{46}$$

where $\hat{S}_{i,m}$ is the statistical significance level associated with $\hat{\psi}_m$ and $\hat{\omega}_m$ is the weight assigned to model *m*, $\sum_m \hat{\omega}_m = 1$. Sala-i-Martin (1997a,b) employs weights determined by the likelihoods of each model as well as employing equal weighting. Second, Sala-i-Martin (1997a,b) proposes examining the percentage of times a variable appears statistically significant with a given sign; a variable whose sign and statistical significance holds across 95% of the different models estimated is regarded as robust. This approach finds that initial income, the investment to GDP ratio and secondary school education are all robust determinants of growth. Sala-i-Martin (1997a,b) extends this analysis to the evaluation of additional variables and finds a number also are robust by his criteria.

While these approaches have the important advantage over extreme bounds analysis of accounting for the informational content of the entire distribution of $\hat{\psi}_m$, the procedures do not have any decision-theoretic or conventional statistical justification. We are unaware of any statistical interpretation to averaged significance levels. Further, little is understood about the statistical properties of these procedures. Hoover and Perez

[42]For further discussion of extreme bounds analysis, see Temple (2000b) and the references therein.

(2004), for example, find that the second Sala-i-Martin procedure has poor size properties, in the sense that "true" growth determinants are still likely to fail to be identified.

Dissatisfaction with extreme bounds analysis and the variants we have described have led some authors to embed the determinants of robust growth regressors in a general model selection context. Hendry and Krolzig (2004) and Hoover and Perez (2004) both employ general-to-specific modeling methodologies generally associated with the research program of David Hendry (cf. Hendry 1995)) to select one version of (41) out of the model space. In both papers, the linear model that is selected out of the space of possible models is one where growth is determined by years an economy is open, the rate of equipment investment, a measure of political instability based on the number of coups and revolutions, a measure of the percentage of the population that is Confucian and a measure of the percentage of the population that is Protestant.

Methodologically, these papers in essence employ algorithms which choose a particular regression model from a space of models through comparisons based on a set of statistical tests. The extent to which one finds this approach appealing is a function of the extent to which one is sympathetic to the general methodological foundations of the Hendry research program; we avoid such an extended evaluation here, but simply note that like other general prescriptions the program remains controversial, especially the extent to which it relies on automatic model selection procedures that do not possess a clear decision-theoretic justification. As such, it is somewhat unclear how to evaluate the output of the procedure in terms of the objectives of a researcher. That being said, the automated procedures Hendry works with have the important virtue that they can facilitate identifying small sets of models that are well supported by available data. Identification of such models is important, for example, in forecasting, where Hendry's procedures appear to have a strong track record.

In our judgment, the most promising current approach to accounting for model uncertainty employs model averaging techniques to construct parameter estimates that formally address the dependence of model-specific estimates on a given model. Examples where model averaging has been applied to cross-country growth data include Brock and Durlauf (2001a), Brock, Durlauf, and West (2003), Doppelhofer, Miller, and

Sala-i-Martin (2004), Fernandez, Ley, and Steel (2001a) and Masanjala and Papageorgiou (2004). The basic idea in this work is to treat the "true" growth model[43] as an unobservable variable. In order to account for this variable, each element *m* in the model space *M* is associated with a posterior model probability $\mu(m|D)$. By Bayes' rule,

$$\mu(m|D) \propto \mu(D|m)\mu(m) \tag{47}$$

where $\mu(D|m)$ is the likelihood of the data given the model and $\mu(m)$ is the prior model probability. These model probabilities are used to eliminate the dependence of parameter analysis on a specific model. For frequentist estimates, averaging is done across the model-specific estimates $\hat{\psi}_m$ to produce an estimate $\hat{\psi}$ via

$$\hat{\psi} = \sum_m \hat{\psi}_m \mu(m|D) \tag{48}$$

whereas for the Bayesian context, the dependence of the posterior probabilility measure of the parameter of interest, $\mu(\psi|D,m)$ on the model choice is eliminated via standard conditional probability arguments, i.e.

$$\mu(\psi|D) = \sum_{m \in M} \mu(\psi|D,m)\mu(m|D) \tag{49}$$

Brock, Durlauf, and West (2003) argue that the strategy of constructing posterior probabilities that are not model-dependent is the appropriate one when the objective of the statistical exercise is to evaluate alternative policy questions such as whether to

[43]In this discussion, we will assume that one of the models in the model space *M* is the correct specification of the growth process. When none of the model specifications is the correct one, this naturally affects the interpretation of the model averaging procedure.

change elements of S_i by one unit. Notice that this approach assumes that the goal of the exercise is to study a parameter, i.e. ψ, not to identify the best growth model.

Model averaging approaches are still quite new in the growth literature, so many questions exist as to how to implement the procedure. One issue concerns the specification of priors on parameters within a model. Brock and Durlauf (2001a), Brock, Durlauf and West (2003), and Doppelhofer, Miller, and Sala-i-Martin (2004) assume a diffuse prior on the model specific coefficients. The advantage of this prior is that, when the errors are normal with known variance, the posterior expected value of ψ, conditional on the data *D* and model m, is the ordinary least squares estimator $\hat{\psi}_m$. The disadvantage of this approach is that since the diffuse prior on the regression parameters is improper, one has to be careful that the posterior model probabilities associated with the prior are interpretable. For this reason, Doppelhofer, Miller, and Sala-i-Martin (2004) eschew reference to their methodology as strictly Bayesian. That being said, so long as the posterior model probabilities include appropriate penalties for model complexity, (and Brock and Durlauf (2001a), Brock, Durlauf, and West (2003), and Doppelhofer, Miller, and Sala-i-Martin (2004) all compute posterior model probabilities using BIC adjusted likelihoods) we do not see any conceptual problem in interpreting this approach as strictly Bayesian. Fernandez, Ley, and Steel (2001a) and Masanjala and Papageorgiou (2004) employ proper priors and therefore avoid such concerns.[44] We are unaware of any evidence that the choice of prior for the within-model regression coefficients is of great importance in terms of empirical inferences for the growth contexts that have been studied; Masanjala and Papageorgiou (2004) in fact compare results using using proper priors with the improper priors we have described and find that the choice of prior is unimportant.

A second unresolved issue concerns the specification of the prior model probabilities $\mu(m)$. In the model averaging literature, the general assumption has been to assign equal prior probabilities to all models in M. This prior may be interpreted as assuming that the prior probability that a given variable appears in the "true" model is .5

[44]Fernandez, Ley, and Steel (2001b) provide a general analysis of proper model specific priors for model averaging exercises.

and that the probability that one variable appears in the model is independent of whether others appear. Doppelhofer, Miller, and Sala-i-Martin (2004) consider modifications of this prior in which the probability that a given variable appears in the true model is $p < .5$; these alternative probabilities are chosen in order to assign greater weight to more parsimonious growth models, i.e. models in which fewer regressors appear.[45]

Brock and Durlauf (2001a) and Brock, Durlauf, and West (2003) argue against the assumption that the probability that one regressor should appear in a growth model is independent of the inclusion of others. The basic problem with priors that assume independence is analogous to the red bus/blue bus problem in discrete choice theory; namely, some regressors are quite similar to others, e.g. alternative measures of trade openness, whereas other regressors are quite disparate, e.g. geography and institutions. Brock, Durlauf, and West (2003) propose a tree structure to organize model uncertainty for linear growth models. First, they argue that growth models suffer from theory uncertainty. Hence, one can identify alternative classes of models based on what growth theories are included. Second, for each specification of a body of theories to be embedded, they argue there is specification uncertainty. A given set of theories requires determining whether the theories interact, whether they are subject to threshold effects or other types of nonlinearity, etc. Third, for each theory and model specification, there is measurement uncertainty. The statement that weather affects growth does not specify the relevant empirical proxies, e.g. the number of sunny days, average temperature, etc. Finally, each choice of theory, specification and measurement is argued to suffer from heterogeneity uncertainty, which means that it is unclear which subsets of countries obey a common linear model. Brock, Durlauf, and West (2003) argue that one should assign priors that account for the interdependences implied by this structure in assigning model probabilities. Appendix 2 follows this approach in organizing growth regressors according to theory.

Doppelhofer, Miller, and Sala-i-Martin (2004) and Fernandez, Ley, and Steel (2001a) employ model averaging methods to identify which growth regressors should be

[45]In our judgment, this presumption is unappealing as our own prior beliefs suggest that the true growth model is likely to contain many distinct factors. One implication of the openendedness of growth theories is that the simultaneous importance of many factors is certainly plausible.

included in linear growth models. These analyses do not distinguish between variables to be included in all regressions and variables whose inclusion determines alternative models; all variables are pooled and all possible combinations are considered. Doppelhofer, Miller, and Sala-i-Martin (2004) working with 31 potential growth determinants, conclude, weighting prior models so that the expected number of included regressors is 7 (this corresponds to a prior probability of variable inclusion of about .25), that four variables have posterior model inclusion probabilities above .9: initial income, fraction of GDP in mining, number of years the economy has been open,[46] and fraction of the population following Confucianism. Working with a universe of 41 potential growth determinants, Fernandez, Ley, and Steel find that, under the assumption that the prior probability that a given variable appears in the correct growth model is .5, four variables have posterior model inclusion probabilities above .9: initial income, fraction of the population following Confucianism, life expectancy, and rate of equipment investment.

Brock and Durlauf (2001a) and Masanjala and Papageorgiou (2004) employ model averaging to study the reason for the poor growth performance of sub-Saharan Africa. Brock and Durlauf (2001a) reexamine Easterly and Levine's (1997a) finding that ethnic heterogeneity helps explain sub-Saharan Africa's growth problems. This reanalysis finds that the Easterly and Levine (1997) claim is robust in the sense that ethnic heterogeneity helps explain why growth in sub-Saharan Africa had stagnated relative to the rest of the world. On the other hand, Brock and Durlauf (2001a) also find that ethnic heterogeneity does not appear to explain growth patterns in the rest of the world. This leads to the unresolved question of why ethnic heterogeneity has uniquely strong growth effects in sub-Sahran Africa. Masanjala and Papageorgiou (2004) conduct a general analysis of the determinants of sub-Saharan African growth versus the world as a whole and conclude that the relevant growth variables for Africa are quite different. In particular, variation in sub-Saharan growth is much more closely associated with the share of the economy made up by primary commodities production. They also find, contrary to Doppelhofer, Miller, and Sala-i-Martin (2004) that the share of mining in the economy is a robust determinant of growth in Africa but not the world as a whole.

[46]Sachs and Warner (1995) use this variable as an index of overall openness of an economy.

Finally, model averaging has been applied by Brock, Durlauf, and West (2003) to analyze the question of how to employ growth regressions to evaluate policy recommendations. Specifically, the paper assesses the question of whether a policymaker should favor a reduction of tariffs for sub-Saharan African countries; the analysis assumes that the policymaker possesses mean/variance preferences with respect to the effects of changes in current policies with a constant tradeoff of mean against standard deviation of 1 to 2. The analysis finds strong support for a tariff reduction in that it concludes that a policymaker with these preferences should support a tariff reduction for any of the countries in sub-Saharan Africa unless the policymaker has a very strong prior that sub-Saharan African countries obey a distinct linear growth process from the rest of the world. In the case where the policymaker has a strong prior that sub-Saharan Africa is "different" from the rest of the world, there is sufficient uncertainty about the relationship between tariffs and growth for these countries that a change in the rates cannot be justified; the strong prior in essence means that the growth experiences of non-African countries have little effect on the precision of estimates of growth behavior that are constructed using data on sub-Saharan African countries in isolation.

ii. parameter heterogeneity

From its earliest stages, the use of linear growth models has generated considerable unease with respect to the statistical foundations of the exercise. Arguably, the data for very different countries cannot be seen as realizations associated with a common data generating process (DGP). For econometricians that have been trained to search for a good approximation to a DGP, the modeling assumptions and procedures of the growth literature can look arbitrary. One expression of this concern is captured in a famous remark in Harberger (1987): "What do Thailand, the Dominican Republic, Zimbabwe, Greece, and Bolivia have in common that merits their being put in the same regression analysis?"

Views differ on the extent to which this objection is fundamental. There is general agreement that, when studying growth, it will be difficult to recover a DGP even if one exists. In particular, the prospects for recovering causal effects are clearly weak. Those

who are only satisfied with the specification and estimation of a structural model, in which parameters are either 'deep' or correspond to precisely defined causal effects within a coherent theoretical framework, will be permanently disappointed.[47] The growth literature must have a less ambitious goal, namely to investigate whether or not particular hypotheses have any support in the data. In practice, growth researchers are looking for patterns and systematic tendencies that can increase our understanding of the growth process, in combination with historical analysis, case studies, and relevant theoretical models. Another key aim of empirical growth research, which is harder than it looks at first sight, is to communicate the degree of support for any patterns identified by the researcher.

The issue of parameter heterogeneity is essentially that raised by Harberger. Why should we expect disparate countries to lie on a common surface? Clearly this criticism could be applied to most empirical work in social science, whether the data points reflect the actions and characteristics of individuals and firms, or the aggregations of their choices that are used in macroeconometrics. What is distinctive about the growth context is not so much the lack of a common surface, as the way in which the sample size limits the scope for addressing the problem. In principle, one response would be to choose a more flexible model that has a stronger chance of being a good approximation to the data. Yet this can be hard, and an inherently fragile procedure, when the sample is rarely greater than 100 observations.

If parameter heterogeneity is present, the consequences are potentially serious, except in a special case. If a slope parameter varies randomly across units, and is distributed independently of the variables in the regression and the disturbances, the coefficient estimate should be an unbiased estimate of the mean of the parameter. The assumption of independence is not one, however, that may be expected in light of the body of growth theories. For example, when estimating the relationship between growth and investment, the marginal effect of investment will almost certainly be correlated with aspects of the economic environment that should also be included in the regression.

[47]Note that this reflects the shortcomings of economic theory as well as those of data and econometric analysis.

The solution to this general problem is to change the specification in a way that allows greater flexibility in estimation. There are many ways of doing this. One approach is to consider more general functional forms than the canonical Solow regression which for comparison purposes we restate as:

$$\gamma_i = k + \beta \log y_{i,0} + \pi_n \log\left(n_i + g + \delta\right) + \pi_K \log s_{K,i} + \pi_H \log s_{H,i} + \varepsilon_i \qquad (50)$$

Liu and Stengos (1999) estimate a semiparametric partially linear version of this model, namely

$$\gamma_i = k + f_\beta\left(\log y_{i,0}\right) + \pi_n \log\left(n_i + g + \delta\right) + \pi_K \log s_{K,i} + f_{\pi_H}\left(\log s_{H,i}\right) + \varepsilon_i \qquad (51)$$

where $f_\beta(\cdot)$ and $f_{\pi_H}(\cdot)$ are arbitrary (except for variance smoothness requirements) functions. One important finding is that the value of $f_\beta\left(\log y_{i,0}\right)$ is only negative when initial per capita income exceeds about \$1800. They also find a threshold effect in secondary school enrollment rates (their empirical proxy for $\log s_{H,i}$) so the variable is only associated with a positive impact on growth if it exceeds about 15%. Banerjee and Duflo (2003) use this same regression strategy to study nonlinearity in the relationship between changes in inequality and growth; their specification estimates a version of (51) where initial income and human capital savings enter linearly (along with some additional non-Solow variables) but with the addition on the right hand side of the function $f_G\left(G_{i,t} - G_{i,t-5}\right)$ where $G_{i,t}$ is the Gini coefficient. Using a panel of 45 countries and 5 year growth averages, their analysis produces an estimate of $f_G(\cdot)$ which has an inverted U shape. One limitation of such studies is that they only allow for nonlinearity for a subset of growth determinants, an assumption that has little theoretical justification and is, from a statistical perspective, ad hoc; of course the approach is more general and less ad hoc then simply assuming linearity as is done in most of the literature.

Durlauf, Kourtellos, and Minkin (2001) extend this approach and estimate a version of the augmented Solow model that allows the parameters for each country to vary as functions of initial income, i.e.

$$\gamma_i = k\left(y_{i,0}\right) + \beta\left(y_{i,0}\right)\log y_{i,0} + \pi_n\left(y_{i,0}\right)\log\left(n_i + \delta + g\right) + \pi_K\left(y_{i,0}\right)\log s_{K,i} + \pi_H\left(y_{i,0}\right)\log s_{H,i} + \varepsilon_i \quad (52)$$

This formulation means that each initial income level defines a distinct Solow regression; as such it shifts the focus away from nonlinearity towards parameter heterogeneity, although the model is of course nonlinear in $y_{i,0}$. This approach reveals considerable parameter heterogeneity especially among the poorer countries. Durlauf, Kourtellos, and Minkin (2001) confirm Liu and Stengos (1999) in finding that $\beta\left(y_{i,0}\right)$ is positive for low $y_{i,0}$ values and negative for higher ones. They also find that $\pi_K\left(y_{i,0}\right)$ fluctuates greatly over the range of $y_{i,0}$ values in their sample. This work is extended in Kourtellos (2003a) who finds parameter dependence on initial literacy and initial life expectancy. The varying coefficient approach is also employed in Mamuneas, Savvides, and Stengos (2004) who analyze annual measures of total factor productivity for 51 countries. They consider a regression model of TFP in which the coefficient on human capital in the regression is allowed to depend on human capital both in isolation and in conjunction with a measure of trade openness (other coefficients are held constant). Constancy of the human capital coefficient is rejected across a range of specifications.

At a minimum, it generally makes sense for empirical researchers to test for neglected parameter heterogeneity, either using interaction terms or by carrying out diagnostic tests. Chesher (1984) showed that White's information matrix test can be used in this context. For the normal linear model with fixed regressors, Hall (1987) showed that, asymptotically, the information matrix test corresponds to a joint test for heteroskedasticity and non-normality. Later in the chapter, we discuss how evidence of heteroskedasticity should sometimes be seen as an indicator of misspecification.

Other authors have attempted to employ panel data to identify parameter heterogeneity without the imposition of a functional relationship between parameters and various observable variables. An important early effort is Canova and Marcet (1995). Defining $s_{i,t}$ as the logarithm of the ratio of a country's per capita income to the time t international aggregate value, Canova and Marcet estimate models of the form

$$s_{i,t} = a_i + \rho_i s_{i,t-1} + \varepsilon_{i,t}. \tag{53}$$

The long-run forecast of $s_{i,t}$ is given by $\frac{a_i}{1-\rho_i}$ with $1-\rho_i$ being the rate of convergence towards that value. Canova and Marcet estimate their model using data on the regions of Europe and on 17 western European countries. Restricting the parameters a_i and ρ_i to be constant across i gives a standard β-convergence test and yields an estimated annual rate of convergence of approximately 2%. On the other hand, allowing for heterogeneity in these parameters produces a "substantial", statistically significant, dispersion of the implied long-run $s_{i,t}$ forecasts. Moreover, those forecasts are positively correlated with $s_{i,0}$, the initial values of $s_{i,t}$, implying a dependence of long-run outcomes on initial conditions contrary to the convergence hypothesis. For the country-level data, differences in initial conditions explain almost half the cross-sectional variation in long-run forecasts; in contrast, the role of standard control variables such as rates of physical and human capital accumulation and government spending shares is minor. The latter finding must be tempered by the fact that the sample variation in these controls is less than that in Barro (1991) or Mankiw, Romer, and Weil (1992), for example.

A similar approach is taken by Maddala and Wu (2000) who consider models of the form

$$\log y_{i,t} = \alpha_i + \rho_i \log y_{i,t-1} + u_{i,t} \tag{54}$$

which is of course very similar to the model analyzed by Marcet and Canova (1995). Employing shrinkage estimators for α_i and ρ_i, they conclude that convergence rates, measured as $\beta_i = -\log \rho_i$ exhibit substantial heterogeneity.

iii. nonlinearity and multiple regimes

In this section we discuss several papers that have attempted to disentangle the roles of heterogeneous structural characteristics and initial conditions in determining growth performance. These studies employ a wide variety of statistical methods in attempting to identify how initial conditions affect growth. Despite this, there is substantial congruence in the conclusions of these papers as these studies each provide evidence of the existence of convergence clubs even after accounting for variation in structural characteristics.

An early contribution to this literature is Durlauf and Johnson (1995) who use classification and regression tree (CART) methods to search for nonlinearities in the growth process as implied by the existence of convergence clubs.[48] The CART procedure identifies subgroups of countries that obey a common linear growth model based on the Solow variables. These subgroups are identified by initial income and literacy, a typical subgroup l is defined by countries whose initial income lies within the interval $\underline{\vartheta}_{l,y} \leq y_{i,0} < \overline{\vartheta}_{l,y}$ and whose literacy rate L_i lies in the interval $\underline{\vartheta}_{l,L} \leq L_i < \overline{\vartheta}_{l,L}$. The number of subgroups and the boundaries for the variable intervals that define them are chosen by an algorithm that trades off model complexity (i.e. the number of subgroups) and goodness of fit. Because the procedure sequentially splits the data into finer and finer subgroups, it gives the data a tree structure.

[48]A detailed discussion of regression tree methods appears in Breiman, Friedman, Olshen and Stone (1984). The technical appendix of Durlauf and Johnson (1995) presents a treatment tailored to the specific question of identifying multiple regimes in growth models. Regression tree methods suffer from the absence of a well-developed asymptotic theory for testing the number of regimes that are present in a data set, but the procedure is consistent in the sense that under relatively weak conditions, if there are a finite number of regimes, as the sample size grows to infinity, the correct model will be revealed.

Durlauf and Johnson (1995) also test the null hypothesis of a common growth regime against the alternative hypothesis of a growth process with multiple regimes in which economies with similar initial conditions tend to converge to one another. Using income per capita and the literacy rate (as a proxy for human capital) to measure the initial level of development and, using the same cross-country data set as Mankiw, Romer, and Weil, Durlauf, and Johnson reject the single regime model required for global convergence. That is, even after controlling for the structural heterogeneity implied by Mankiw, Romer, and Weil's augmented version of the Solow model, there is a role for initial conditions in explaining variation in cross-country growth behavior.

Durlauf and Johnson's (1995) findings of multiple convergence clubs appear to be reinforced by subsequent research. Papageorgiou and Masanjala (2004) note that one possible source for Durlauf and Johnson's findings may occur due to the misspecification of the aggregate production function. As observed in Section II, the linear representation of the Solow model represents an approximation around the steady-state when the aggregate production function is Cobb-Douglas. Papageorgiou and Masanjala estimate a version of the Solow model based on a constant elasticity of substitution (CES) production function rather than the Cobb-Douglas, following findings in Duffy and Papageorgiou (2000). They then examine the question of whether or not Durlauf and Johnson's multiple regimes remain under the CES specification. Using Hansen's (2000) approach to sample splitting and threshold estimation, they find statistically significant evidence of thresholds in the data. The sample splits they estimate divide the data in four distinct growth regimes and are broadly consistent with those found by Durlauf and Johnson.[49]

[49]Motivated by the debate over trade openness and growth, Papageorgiou (2002) applies Hansen's method to the Durlauf and Johnson data with the trade share added to the set of variables on which sample splits may occur. He finds that this variable divides the middle-income countries into high and low growth groups obeying different growth processes; however openness does not appear to matter for high and low income countries. This suggests the importance of further work on which variables are most appropriate in characterizing threshold effects. Using the regression tree approach with a large collection of candidate split variables, Johnson and Takeyama (2001) find evidence of thresholds in US state economic growth behavior defined by variables likely to be proxies for capital/labor ratio, agglomeration effects, and communication effects.

These findings are extended in recent work due to Tan (2004) who employs a procedure known as GUIDE (generalized, unbiased interaction detection and estimation) to identify subgroups of countries which obey a common growth model.[50] Relative to CART, the GUIDE algorithm has two advantages: 1) the algorithm explicitly looks for interactions between explanatory variables when identifying splits and 2) some within model testing supplements the penalties for model complexity and thereby reduces the tendency for CART procedures to produce an excessive number of splits in finite samples. Tan (2004) finds strong evidence that measures of institutional quality and ethnic fractionalization define convergence clubs across a wide range of countries. He also finds weaker evidence that geography distinguishes the growth process for sub-Saharan Africa from the rest of the world.

Further research has corroborated the evidence of multiple regimes using alternative statistical methods. One approach that has proven useful is based on projection pursuit methods[51]. Desdoigts (1999) uses these methods in an attempt to separate the roles of microeconomic heterogeneity and initial conditions in the growth experiences of a group of countries and identifies groups of countries with relatively homogeneous growth experiences based on data about the characteristics and initial conditions of each country. The idea of projection pursuit is to find the orthogonal projections of the data into low dimensional spaces that best display some interesting feature of the data. A well-known special case of projection pursuit is principal components analysis. In principal components analysis, one takes only as many components as are necessary to account for "most" of the variation in the data. Similarly, in projection pursuit one should only consider as many dimensions as needed to account for "most" of the clustering in the data.

Desdoigts finds several interesting clusters. The first is the OECD countries. The two projections identifying this cluster put most of their weight on the primary and secondary school enrollment rates, the 1960 income gap and the rate of growth in the labor force. The prominence of variables that Desdoigts argues are proxies for initial

[50]GUIDE originates in Loh (2002).
[51]Projection pursuit is developed in Friedman and Tukey (1974) and Friedman (1987). Appendix A of Desdoigts (1999) provides a useful primer.

conditions among those defining the projections leads him to conclude that initial conditions are more important in defining this cluster than are other country characteristics. Reapplication of the clustering method to the remaining (non-OECD) countries yields three sub-clusters that can be described as Africa, Southeast Asia, and Latin America. Here the projections put most weight on government consumption, the secondary school enrollment rate and investment in electrical machinery and transportation equipment. Most of these variables are argued to proxy for structural characteristics of the economies, suggesting that they, rather than initial conditions, are responsible for the differences in growth experiences across the three geographic sub-clusters. Nevertheless, this approach relies on the judgment of the researcher in determining which variables proxy for initial conditions and which proxy for structural characteristics.

Further evidence of the utility of projection pursuit methods may be found in Kourtellos (2003b). Unlike Desdoigts, Kourtellos (2003b) uses projection pursuit to construct models of the growth process. Formally, he estimates models of the form

$$\gamma_i = \sum_{l=1}^{L} f_l \left(y_{i,0}\beta_l + X_i\psi_l + Z_i\pi_l \right) + \varepsilon_i \tag{55}$$

Each element in the summation represents a distinct projection. Kourtellos uncovers evidence of two steady-states, one for low initial income and low initial human capital countries.

A third approach to multiple regimes is employed by Bloom, Canning, and Sevilla (2003) based on the observation that if long-run outcomes are determined by fundamental forces alone, the relationship between exogenous variables and income levels ought to be unique. If initial conditions play a role there will be multiple relationships – one for each basin of attraction defined by initial conditions. If there are two (stochastic) steady states, and large shocks are sufficiently infrequent, [52] the system will, under suitable regularity

[52]The assumed rarity of large shocks implies that movements between basins of attraction of each of the steady states are sufficiently infrequent that they can be ignored in

conditions, exhibit an invariant probability measure that can be described by a "reduced form" model in which the long-run behavior of $\log y_{i,t}$ depends only on the exogenous variables, m_i, such as

$$\log y_{i,t} = \log y_1^*(m_i) + u_{1,i,t} \text{ with probability } p(m_i) \tag{56}$$

and

$$\log y_{i,t} = \log y_2^*(m_i) + u_{2,i,t} \text{ with probability } 1 - p(m_i) \tag{57}$$

where $u_{1,i,t}$ and $u_{2,i,t}$ are independent, zero-mean deviations from the steady-state log means $\log y_1^*(m_i)$ and $\log y_2^*(m_i)$ respectively, and $p(m_i)$ is the probability that country i is in the basin of attraction of the first of the two steady states. From the perspective of the econometrician, $\log y_{i,t}$ thus obeys a mixture process. The two steady states associated with (56) and (57) are possibly interpretable as a low-income regime or poverty trap and as a high-income or perpetual growth regime respectively. Bloom, Canning and Sevilla estimate a linear version of this model using 1985 income data from 152 countries with the absolute value of the latitude of the (approximate) center of each country as the fundamental exogenous variable. They are able to reject the null hypothesis of a single regime model in favor of the alternative of a model with two regimes – a high-level (manufacturing and services) steady state in which income is independent of latitude and a low-level (agricultural) steady-state in which income depends on latitude (presumably through its influence on climate). In addition, the probability of being in the high-level steady state is found to rise with latitude.

A final approach to multiple regimes is due to Canova (2004) who introduces a procedure for panel data that estimates the number of groups and the assignment of countries or regions to these groups, drawing on Bayesian ideas. This approach has the

---

estimation. This assumption is consistent with, for example, Bianchi's (1997) finding of very little mobility in the cross-country income distribution.

important feature that it allows for parameter heterogeneity across-countries within a given subgroup. The researcher can order the countries or regions by various criteria (for example, output per capita in the pre-sample period) and the estimation procedure then chooses break points and group membership in such a way that the predictive ability of the overall model is maximized. This approach is applied to autoregressive models of per capita output as in eq. (54) above.

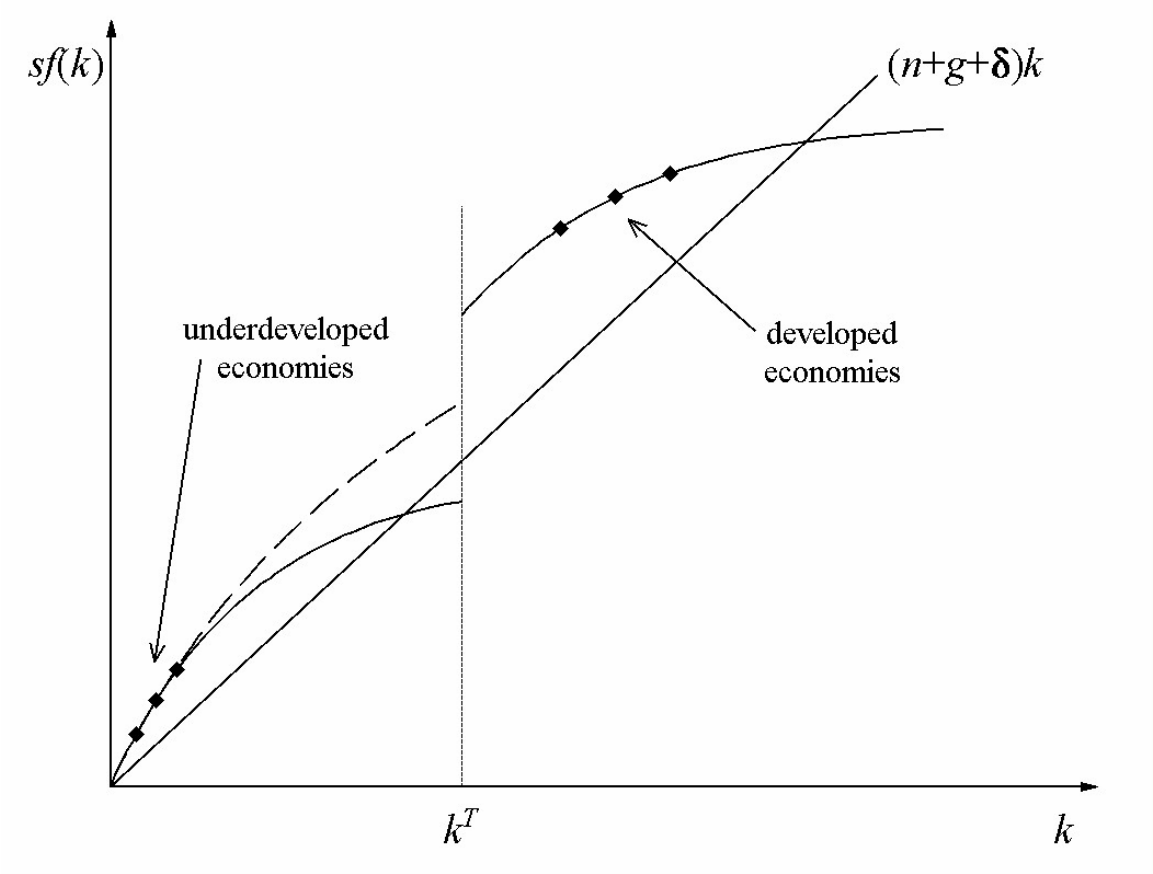
Using data on per capita income data in the regions of Europe, Canova (2004) finds that ordering the data by initial income maximizes the marginal likelihood of the model and breaks the data into 4 clusters. The estimated mean steady-states for each group are significantly different from each other implying that the groups are convergence clubs. The differences in the means are also economically important with the lowest and highest being 45% and 115% of the overall average respectively. Canova finds little across-group mobility especially among those regions that are initially poor. Using data on per capita income in the OECD countries, two clusters are found and, again, initial per capita income is the preferred ordering variable. The estimated model parameters imply an "economically large" long-run difference in the average incomes of countries in the two groups with little mobility between them.

In assessing these analyses, it is important to recognize an identification problem in attempting to link evidence of multiple growth regimes to particular theoretical growth models. As argued in Durlauf and Johnson (1995), this identification problem relates to whether evidence of multiple regimes represents evidence of multiple steady-states as opposed to nonlinearity in the growth process.

To see why this is so, suppose that one has identified two sets of countries that obey separate growth regimes with regime membership determined by a country's initial capital stock, i.e. there exists a capital threshold k^T that divides the two groups of countries. An example of this can be seen in Figure 8. Clearly, the two sets of countries do not obey a common linear model but it is not clear whether or not multiple steady-states exist. The output behavior of low capital countries is compatible with either the solid or dashed curve in the lower part of the figure, but only the solid curve produces multiple steady-states. The identification problem stems from the fact that one does not have observations that allow one to distinguish differences in the long-run behavior of

countries that start with capital stocks in the vicinity of k^T. This argument does not depend on growth regimes determined by the capital stock but it does depend on whether or not the variable or variables that define the regimes are growing over time, as would occur for initial income or initial literacy. For growing variables, the possibility exists that countries currently associated with low levels of the variable will in the future exhibit behaviors that are similar to those countries which are currently associated with high levels of the variables.

Figure 8: Nonlinearity Versus Multiple Steady-States



How might evidence of multiple steady-states be achieved? One possibility is via the use of structural models in empirical analysis. While this has not been done econometrically, Graham and Temple (2003) follow this strategy and calibrate a two-sector general equilibrium model with increasing returns to scale in nonagricultural production. Their empirically motivated choice of calibration parameters produces a model which implies that some countries are in a low-output equilibrium. Another possibility is to exploit time series variation in a single country to identify the presence of jumps from one equilibrium or steady state to another.

VI. Econometric issues I: alternative data structures

Our discussion of growth econometrics now shifts from general issues of hypothesis testing and model specification to explore specific econometric issues that arise in the estimation of growth models. This section reviews econometric issues that arise for the different types of data structures that appear in growth analyses. By data structures, we refer to features such as whether the data are observed in cross-section, time series, or panel as well as to whether particular data series are conceptualized as endogenous or exogenous. At the risk of stating the obvious, choices of method involve significant trade-offs, which depend partly on statistical considerations and partly on the economic context. This means that attempts at universal prescriptions are misguided, and we will try to show the desirability of matching techniques to the economic question at hand.

One example, to be discussed further below, would be the choice between panel data methods and the estimation of separate time series regressions for each country. The use of panel data is likely to increase efficiency and allow richer models to be estimated, but at the expense of potentially serious biases if the parameter homogeneity assumptions are incorrect. This trade-off between robustness and efficiency is another running theme of our survey. The scientific solution would be to base the choice of estimation method on a context-specific loss function, but this is clearly a difficult task, and in practice more subjective decisions are involved.

This section has four main elements. Section VI.i examines econometric issues that arise in the use of time series data to study growth, emphasizing some of the drawbacks of this approach. Section VI.ii discusses the many issues that arise when panel data are employed, an increasingly popular approach to growth questions. We consider the estimation of dynamic models in the presence of fixed effects, and alternatives to standard procedures. Section VI.iii describes another increasingly popular approach, namely the use of "event studies" to analyze growth behavior, based on studying responses to major shocks such as policy reforms. Section VI.iv examines endogeneity and the use of instrumental variables. We argue that the use of instrumental variables in growth contexts is more problematic than is often appreciated and suggest

the importance of combining instrumental variable choice with a systematic approach to model selection.

i. time series approaches

At first glance, the most natural way to understand growth would be to examine time series data for each country in isolation. As we saw previously, growth varies substantially over time, and countries experience distinct events that contribute to this variation, such as changes in government and in economic policy.

In practice a time series approach runs into substantial difficulties. One key constraint is the available data. For many developing countries, some of the most important data are only available on an annual basis, with limited coverage before the 1960s. Moreover, the listing of annual data in widely used sources and online databases can be misleading, because some key variables are measured less frequently. For example, population figures are often based primarily on census data, while measures of average educational attainment are often constructed by interpolating between census observations using school enrollments. When examining published data, it is not always clear where this kind of interpolation has been used. The true extent of information in the time series variation may be less than appears at first glance, and conventional standard errors on parameter estimates will be misleading when interpolated data are used.

Even where reliable data are available, some key growth determinants display relatively little time variation, a point that has been emphasized by Easterly et al (1993), Easterly (2001) and Pritchett (2000a). There do exist other variables that appear to show significant variation, but this variation may not correspond to the concept the researcher has in mind. An example would be political stability. Since Barro (1991), researchers have sometimes used the incidence of political revolutions and coups as a measure of political instability. The interpretation of such an index clearly varies depending on the length of the time period used to construct it. If the hypothesis of interest relates to underlying political uncertainty (say, the ex-ante probability of a transfer of power) then the observations on political instability would need to be averaged over a long time

period. The variation in political instability at shorter horizons only casts light on a different hypothesis, namely the direct impact of revolutions and coups.

There are other significant problems with the time series approach. The hypotheses of most interest to growth theorists are mainly about the evolution of potential output, not deviations from potential output such as business cycles and output collapses. Since measured output is a noisy indicator of potential output, it is easy for the econometric modeling of a growth process to be contaminated by business cycle dynamics. A simple way to illustrate this would be to consider what happens if measured log output is equal to the log of potential output plus a random error. If log output is trend stationary, this is a classical measurement error problem. When lags of output or the growth rate are used as explanatory variables, the parameter estimates will be inconsistent.

Such problems are likely to be even more serious in developing countries, where large slumps or crises are not uncommon, and output may deviate for long periods from any previous structural trend (Pritchett (2000a)). We have already seen the extent to which output behaves very differently in developing countries compared to OECD members, and a major collapse in output is not a rare event. There may be no underlying trend in the sense commonly understood, and conventional time series methods should be applied with caution. Some techniques that are widely used in the literature on business cycles in developed countries, such as the Hodrick-Prescott filter, will often be inappropriate in the context of developing countries.

The problem of short-run output instability extends further. It is easy to construct examples where the difference between observed output and potential output is correlated with variables that move up and down at high frequencies, with inflation being one obvious example. This means that time series studies of inflation and growth based on observed output will find it hard to isolate reliably an effect of inflation on potential output; for further discussion see Temple (2000a). When considerations like these are combined with the paucity of the available data, it appears a hard task to learn about long-term growth using time series regressions, especially when developing countries are the main focus of interest.

Nevertheless, despite these problems, there are some hypotheses for which time series variation can be informative. We have already seen the gains from time series approaches to convergence issues. Jones (1995) and Kocherlakota and Yi (1997) show how time series models can be used to discriminate between different growth theories. To take the simplest example, the AK model of growth predicts that the growth rate will be a function of the share of investment in GDP. Jones points out that investment rates have trended upwards in many OECD countries, with no corresponding increase in growth rates. Although this might be explained by offsetting changes in other growth determinants, it does provide evidence against simple versions of the AK model.

Jones (1995) and Kocherlakota and Yi (1997) develop a statistical test of endogenous growth models based on regressing growth on lagged growth and a lagged policy variable (or the lagged investment rate, as in Jones). Exogenous growth models predict that the coefficients on the lagged policy variable should sum to zero, indicating no long-run effect of permanent changes in this variable on the growth rate. In contrast, some endogenous growth models imply that the sum of coefficients should be non-zero. A simple time series regression then provides a direct test of the predictions of these models. More formally, as in Jones (1995), for a given country *i* one can investigate a dynamic relationship for the growth rate $\gamma_{i,t}$ where

$$\gamma_{i,t} = A(L)\gamma_{i,t-1} + B(L)z_{i,t} + \varepsilon_{i,t} \tag{58}$$

.

where z is the policy variable or growth determinant of interest, and $A(L)$ and $B(L)$ are lag polynomials assumed to be compatible with stationarity. The hypothesis of interest is whether $B(1) \neq 0$. If the sum of the coefficients in the lag polynomial $B(L)$ is significantly different from zero, this implies that a permanent change in the variable *z* will affect the growth rate indefinitely. As Jones (1995) explicitly discusses, this test is best seen as indicating whether a policy change affects growth over a long horizon, rather than firmly identifying or rejecting the presence of a long-run growth effect in the theoretical sense of that term. The theoretical conditions under which policy variables

affect the long-run growth rate are remarkably strict, and many endogenous growth models are best seen as new theories of potentially sizeable level effects.[53]

This approach is closely related to Granger-causality testing, where the hypothesis of interest would be the explanatory power of lags of $Z_{i,t}$ for $\gamma_{i,t}$ conditional on lagged values of $\gamma_{i,t}$. Blomstrom, Lipsey, and Zejan (1996) carry out Granger-causality tests for investment and growth using panel data with five-year subperiods. They find strong evidence that lagged growth rates have explanatory power for investment rates, but much weaker evidence for causality in the more conventional direction from investment to growth. Hence, the partial correlation between growth and investment found in many cross-section studies may not reflect a causal effect of investment. In a similar vein, Campos and Nugent (2002) find that, once Granger-causality tests are applied, the evidence that political instability affects growth may be weaker than usually believed.

The motivation for these two studies, and others like them, is that evidence of temporal precedence helps to build a case that one variable is influenced by another. When this idea is extended to panels, an underlying assumption is that timing patterns and effects will be similar across units (countries or regions). Potential heterogeneity has sometimes been acknowledged, as in the observation of Campos and Nugent (2002) that their results are heavily influenced by the African countries in the sample. The potential importance of these factors is also established in Binder and Brock (2004) who, by using panel methods to allow for heterogeneity in country-specific dynamics, find feedbacks from investment to growth beyond those that appear in Blomstrom, Lipsey, and Zejan (1996).

A second issue is more technical. Since testing for Granger-causality using panel data requires a dynamic model, the use of a standard fixed effects (within groups) estimator is likely to be inappropriate when individual effects are present. We discuss this further in section VI below. One potential solution is the use of instrumental variable procedures, as in Campos and Nugent (2002). In the context of investment and growth, a comprehensive examination of the associated econometric issues has been carried out by

[53]See Temple (2003) for more discussion of this point and the long-run implications of different growth models.

Bond, Leblebicioglu, and Schiantarelli (2004). Their work shows that these issues are more than technicalities: unlike Blomstrom, Lipsey, and Zejan (1996), they find strong evidence that investment has a causal effect on growth.

A familiar objection to the more ambitious interpretations of Granger-causality is that much economic behavior is forward-looking (see for example Klenow and Rodriguez-Clare (1997b)). The movements of stock markets are one instance where temporal sequences can be misleading about causality. Similarly, when entrepreneurs or governments invest heavily in infrastructure projects, or when unusually high inflows of foreign direct investment are observed, the fact that such investments precede strong growth does not establish a causal effect.

ii. panel data

As we emphasized above, the prospects for reliable generalizations in empirical growth research are often constrained by the limited number of countries available. This constraint makes parameter estimates imprecise, and also limits the extent to which researchers can apply more sophisticated methods, such as semiparametric estimators.

A natural response to this constraint is to use the within-country variation to multiply the number of observations. Using different episodes within the same country is ultimately the only practical substitute for somehow increasing the number of countries. To the extent that important variables change over time, this appears the most promising way to sidestep many of the problems that face growth researchers. Moreover, as the years pass and more data become available, the prospects for informative work of this kind can only improve.

We first discuss the implementation and advantages of panel data estimators in more detail, and then some of the technical issues that arise in the context of growth. Perhaps not surprisingly, these methods introduce a set of problems of their own, and should not be regarded as a panacea. Too often, panel data results are interpreted without sufficient care and risk leading researchers astray. In particular, we highlight the care needed in interpreting estimates based on fixed effects.

We will use T to denote the number of time series observations in a panel of N countries or regions. At first sight, T should be relatively high in this context, because of the availability of annual data. But the concerns about time series analysis raised above continue to apply. Important variables are either measured at infrequent intervals, or show little year-to-year variation that can be used to identify their effects. Moreover, variation in growth rates at annual frequencies may give very misleading answers about the longer-term growth process. For this reason, most panel data studies in the growth field have averaged data over five or ten year periods. Given the lack of data before 1960, this implies that growth panels not only have relatively few cross-sectional units (the number of countries employed is often between 50 and 100) but also very low values of T, often 5 or 6 at most.[54]

Most empirical growth models estimated using panel data are based on the hypothesis of conditional convergence, namely that countries converge to parallel equilibrium growth paths, the levels of which are a function of a few variables. A corollary is that an equation for growth (essentially the first difference of log output) should contain some dynamics in lagged output. In this case, the growth equation can be rewritten as a dynamic panel data model in which current output is regressed on controls and lagged output, as in Islam (1995). In statistical terms this is the same model, the only difference of interpretation being that the coefficient on initial output (originally β) is now $1+\beta$:

$$\log y_{i,t} = (1+\beta)\log y_{i,t-1} + \psi X_{i,t} + \pi Z_{i,t} + \alpha_i + \mu_t + \varepsilon_{i,t} \qquad (59)$$

This regression is a general panel analog to the cross-section regression (18). In this formulation, α_i is a country-specific effect and μ_t is a time-specific effect. The inclusion of time-specific effects is important in the growth context, not least because the means of the log output series will typically increase over time, given productivity growth at the world level.

[54]This is true of the many published studies that have used version 5.6 of the Penn World Tables. Now that more recent data are available, there is more scope for estimating panels with a longer time dimension.

Inclusion of a country-specific effect allows permanent differences in the level of income between countries that are not captured by $X_{i,t}$ or $Z_{i,t}$. In principle, one can also allow the parameters $1+\beta$, ψ, and π to differ across *i*; Lee, Pesaran, and Smith (1997,1998) do this for the coefficients for $\log y_{i,t-1}$ and a linear time trend (the latter allowing for steady-state differences in the rate of technological change, corresponding to non-parallel growth paths in the steady state).

The vast majority of panel data growth studies use a fixed effects (within-group) estimator rather than a random effects estimator. Standard random effects estimators require that the individual effects α_i are distributed independently of the explanatory variables, and this requirement is clearly violated for a dynamic panel such as (59) by construction, given the dependence of $\log y_{i,t}$ on α_i.

Given the popularity of fixed effects estimators, it is important to understand how these estimators work. In a fixed effects regression there is a full set of country-specific intercepts, one for each country, and inference proceeds conditional on the particular countries observed (a natural choice in this context). Identification of the slope parameters, usually constrained to be the same across countries, relies on variation over time within each country. The "between" variation, namely the variation across countries in the long-run averages of the variables, is not used.

The key strength of this method, familiar from the microeconometric literature, is the ability to address one form of unobserved heterogeneity: any omitted variables that are constant over time will not bias the estimates, even if the omitted variables are correlated with the explanatory variables. Intuitively, the country-specific intercepts can be seen as picking up the combined effects of all such variables. This is the usual motivation for using fixed effects in the growth context, especially in estimating conditional convergence regressions, as is further discussed in Islam (1995), Caselli Esquivel, and Lefort (1996) and Temple (1999). A particular motivation for the use of fixed effects arises from the Mankiw, Romer, and Weil (1992) implementation of the Solow model. As discussion in Section III, their version of the model implies that one determinant of the level of the steady-state growth path is the initial level of efficiency ($A_{i,0}$) and cross-section heterogeneity in it should usually be regarded as unobservable,

cf. eq. (15). Islam (1995) explicitly develops a specification in which this term is treated as a fixed effect, while world growth and common shocks are incorporated using time-specific effects.

The use of panel data methods to address unobserved heterogeneity can bring substantial gains in robustness, but is not without costs. The fixed-effects identification strategy cannot be applied in all contexts. Sometimes a variable of interest is measured at only one point in time. Even where variables are measured at more frequent intervals, some are highly persistent, in which case the within-country variation is unlikely to be informative. At one extreme, some explanatory variables of interest are essentially fixed factors, like geographic characteristics or ethnolinguistic diversity. Here the only available variation is "between-country", and empirical work will have to be based on cross-sections or pooled cross-section time-series. Alternatively a two-stage hybrid of these methods can be used, in which a panel data estimator is used to obtain estimates of the fixed effects, which are then explicitly modeled in a second stage as in Hoeffler (2002). As we discuss further below, an important direction for future panel data work may be the analysis of the information content of country-specific effects.

A common failing of panel data studies based on within-country variation is that researchers do not pay enough attention to the dynamics of adjustment. There are many panel data papers on human capital and growth that test only whether a change in school enrollment or years of schooling has an immediate effect on aggregate productivity, which seems an implausible hypothesis. Another example, given by Pritchett (2000a), is the use of panels to study inequality and growth. All too often, changes in the distribution of income are implicitly expected to have an immediate impact on growth. Yet many of the relevant theoretical papers highlight long-run effects, and there is a strong presumption that much of the short-run variation in measures of inequality is due to measurement error. In these circumstances, it is hard to see how the available within-country variation can shed much useful light.

There is also a more general problem. Since the fixed effects estimator ignores the between-country variation, the reduction in bias typically comes at the expense of higher standard errors. Another reason for imprecision is that either of the devices used to eliminate the country-specific intercepts – the within-groups transformation or first-

differencing – will tend to exacerbate the effect of measurement error.[55] As a result, it is common for researchers using panel data models with fixed effects, especially in the context of small *T*, to obtain imprecise sets of parameter estimates.

Given the potentially unattractive trade-off between robustness and efficiency, Barro (1997), Temple (1999), Pritchett (2000a) and Wacziarg (2002) all argue that the use of fixed effects in empirical growth models has to be approached with care. The price of eliminating the misleading component of the between variation – namely, the variation due to unobserved heterogeneity – is that all the between variation is lost.

There are alternative ways to reveal this point, but consider the random effects GLS estimator of the slope parameters, which will be more efficient than the within-country estimator for small *T* when the random-effects assumptions are appropriate. This GLS estimator can be written as a matrix-weighted average of the within-country estimator and the between-country estimator, which is based on averaging the data over time and then estimating a simple cross-section regression by OLS.[56] The weights on the two sets of parameter estimates are the inverses of their respective variances. The corollary of high standard errors using within-country estimation, indicating that the within-country variation is relatively uninformative, is that random effects estimates based on a panel of five-yearly averages are very similar to OLS estimates based on thirty-year averages (Wacziarg (2002)). Informally, the random effects estimator sees the between-country variation as offering the greatest scope for identifying the parameters.[57]

This should not be surprising: growth episodes within countries inevitably look a great deal more alike than growth episodes across countries, and therefore offer less identifying variation. Restricting the analysis to the within variation eliminates one source of bias, but immediately makes it harder to identify growth effects with any degree of precision. This general problem is discussed in Pritchett (2000a). Many of the

[55]See Arellano (2003, p. 47-51) for a more formal treatment of this issue.
[56]This result holds for the GLS estimator of the random effects model. In practice, since the true variance components are unknown, feasible GLS must be used.
[57]Of course, this does not imply that the random effects estimator is the best choice; as we have seen, the underlying assumptions for consistency of the estimator are necessarily invalid for a dynamic panel. Instead, our discussion is intended to draw attention to the trade-off between bias and efficiency in deciding whether or not to use fixed-effects estimation.

explanatory variables currently used in growth research are either highly stable over time, or tending to trend in one direction. Educational attainment is an obvious example. Without useful identifying variation in the time series data, the within-country approach is in trouble. Moreover, growth is quite volatile at short horizons. It will typically be hard to explain this variation using predictors that show little variation over time, or that are measured with substantial errors. The result has been a number of panel data studies suggesting that a given variable "does not matter" when a more accurate interpretation is that its effect cannot be identified using the data at hand.

Some of these problems suggest a natural alternative to the within-country estimator, which is to devote more attention to modeling the heterogeneity, rather than treating it as unobserved (Temple (1999)). To put this differently, current panel data methods treat the individual effects as nuisance parameters. As argued by Durlauf and Quah (1999) this is clearly inappropriate in the growth context. The individual effects are of fundamental interest to growth economists because they appear to be a key source of persistent income differences. This suggests that more attention should be given to modeling the heterogeneity rather than finding ways to eliminate its effects.[58]

Depending on the sources of heterogeneity, even simple recommendations, such as including a complete set of regional dummies, can help to alleviate the biases associated with omitted variables. More than a decade of growth research has identified a host of fixed factors that could be used to substitute for country-specific intercepts. A growth model that includes these variables can still exploit the panel structure of the data, and overall this approach has clear advantages in both statistical and economic terms. It means that the between variation is retained, rather than entirely thrown away, while the explicit modeling of the country-specific effects is directly informative about the sources of persistent income and growth differences.

In practice, the literature has focused on another aspect of using panel data estimators to investigate growth. Nickell (1981) showed that within-groups estimates of a

[58]Note that fixed-effects estimators could retain a useful role, because it would be natural to compare their parameter estimates with those obtained using a specific model for the heterogeneity. Where the estimates of common parameters, such as the coefficient on the lagged dependent variable, are different across the two methods, this could indicate the chosen model for the heterogeneity is misspecified.

dynamic panel data model can be badly biased for small *T*, even as *N* goes to infinity. The direction of this bias is such that, in a growth model, output appears less persistent than it should (the estimate of β is too low) and the rate of conditional convergence will be overestimated.

In other areas of economics, it has become increasingly common to avoid the within-groups estimator when estimating dynamic models. The most widely-used alternative strategy is to difference the model to eliminate the fixed effects, and then use two stage least squares or GMM to address the correlation between the differenced lagged dependent variable and the induced MA(1) error term. To see the need for instrumental variable procedures, first-difference (59) to obtain

$$\Delta \log y_{i,t} = (1+\beta)\Delta \log y_{i,t-1} + \Delta X_{i,t}\psi + \Delta Z_{i,t}\pi + \Delta \mu_i + \varepsilon_{i,t} - \varepsilon_{i,t-1} \qquad (60)$$

and note that (absent an unlikely error structure) the $\log y_{i,t-1}$ component of $\Delta \log y_{i,t-1}$ will be correlated with the ε_{it-1} component of the new composite error term, as is clearly seen by considering equation (59) lagged one period. Hence, at least one of the explanatory variables in the first-differenced equation will be correlated with the disturbances, and instrumental variable procedures are required.

Arellano and Bond (1991), building on work by Holtz-Eakin, Newey, and Rosen (1988), developed the GMM approach to dynamic panels in detail, including methods suitable for unbalanced panels and specification tests. Caselli, Esquivel, and Lefort (1996) applied their estimator in the growth context and, as discussed above, this approach yielded a much faster rate of conditional convergence than found in cross-section studies.

The GMM approach is typically based on using lagged levels of the series as instruments for lagged first differences. If the error terms in the levels equation (ε_{it}) are serially uncorrelated then $\Delta \log y_{i,t-1}$ can be instrumented using $\log y_{i,t-2}$ and earlier lagged levels (where available). This corresponds to a set of moment conditions that can be used to estimate the first-differenced equation by GMM. Bond (2002) provides an accessible introduction to this approach.

As an empirical strategy for growth research, this has some appeal, because it could alleviate biases due to measurement error and endogenous explanatory variables. In practice, many researchers are skeptical that lags are suitable instruments. It is easy to see that a variable such as educational attainment may influence output with a considerable delay, so that the exclusion of lags from the growth equation can look arbitrary. More generally, the GMM approach relies on a lack of serial correlation in the error terms of the growth equation (before differencing). Although this assumption can be tested using the methods developed in Arellano and Bond (1991), and can also be relaxed by an appropriate choice of instruments, it is nevertheless restrictive in some contexts.

Another concern is that the explanatory variables may be highly persistent, as is clearly true of output. Lagged levels can then be weak instruments for first differences, and the GMM panel data estimator is likely to be severely biased in short panels. Bond, Hoeffler, and Temple (2001) illustrate this point by comparing the Caselli, Esquivel, and Lefort (1996) estimates of the coefficient on lagged output with OLS and within-group estimates. Since the OLS and within-group estimates of β are biased in opposing directions then, leaving aside sampling variability and small-sample considerations, a consistent parameter estimate should lie between these two extremes (see Nerlove (1999,2000)). Formally, when the explanatory variables other than lagged output are strictly exogenous, we have

$$p\lim \hat{\beta}_{WG} < p\lim \hat{\beta} < p\lim \hat{\beta}_{OLS} \tag{61}$$

where $\hat{\beta}$ is a consistent parameter estimate, $\hat{\beta}_{WG}$ is the within-groups estimate and $\hat{\beta}_{OLS}$ is the estimate from a straightforward pooled OLS regression. For the data set and model used by Caselli, Esquivel and Lefort, this large-sample prediction is not valid, which raises a question mark over the reliability of the first-differenced GMM estimates.

One device that can be informative in short panels is to make more restrictive assumptions about the initial conditions. If the observations at the start of the sample are distributed in a way that is representative of steady-state behavior, in a sense that can be made more precise, efficiency gains are possible. Assumptions about the initial

conditions can be used to derive a "system" GMM estimator, of the form developed and studied by Arellano and Bover (1995) and Blundell and Bond (1998), and also discussed in Ahn and Schmidt (1995) and Hahn (1999). In this estimator, not only are lagged levels used as instruments for first differences, but lagged first differences are used as instruments for levels, which corresponds to an extra set of moment conditions.

There is some Monte Carlo evidence (Blundell and Bond (1998)) that this estimator is more robust than the Arellano-Bond method in the presence of highly persistent series. As also shown by Blundell and Bond (1998), the necessary assumptions can be seen in terms of an extra restriction, namely that the deviations of the initial values of $\log y_{i,t}$ from their long-run values are not systematically related to the individual effects.[59] For simplicity, we focus on the case where there are no explanatory variables other than lagged output. The required assumption on the initial conditions is that, for all $i = 1, \ldots, N$ we have

$$E\left[\left(\log y_{i,1} - \bar{y}_i\right)\alpha_i\right] = 0 \tag{62}$$

where the $\bar{y}_i$ are the long-run values of the $\log y_{i,t}$ series and are therefore functions of the individual effects α_i and the autoregressive parameter β. This assumption on the initial conditions ensures that

$$E\left[\Delta \log y_{i,2}\alpha_i\right] = 0 \tag{63}$$

and this together with the mild assumption that the changes in the errors are uncorrelated with the individual effects, i.e.

$$E\left[\Delta \varepsilon_{i,t}\alpha_i\right] = 0 \tag{64}$$

[59]Note that the long-run values of log output are evolving over time when time-specific effects are included in the model.

implies $T-2$ extra moment conditions of the form

$$E\left[\Delta \log y_{i,t-1}(\alpha_i + \varepsilon_{i,t})\right] = 0 \text{ for } i = 1,\ldots,N \text{ and } t = 3,\ 4,\ldots,T \qquad (65)$$

Intuitively, as is clear from the new moment conditions, the extra assumptions ensure that the lagged first difference of the dependent variable is a valid instrument for untransformed equations in levels since it is uncorrelated with the composite error term in the levels equation. These extra moment conditions can then be combined with the more conventional conditions used in the Arellano-Bond method. This builds in some insurance against weak identification, because if the series are persistent and lagged levels are weak instruments for first differences, it may still be the case that lagged first differences have some explanatory power for levels.[60]

In principle, the validity of the restrictions on the initial conditions can be tested using the incremental Sargan statistic (or C statistic) associated with the additional moment conditions. Yet the validity of the restriction should arguably be evaluated in wider terms, based on some knowledge of the historical forces giving rise to the observed initial conditions. This point – that key statistical assumptions should not always be evaluated only in statistical terms – is one that we will return to later.

Alternatives to GMM have been proposed. Kiviet (1995,1999) derives an analytical approximation to the Nickell bias that can be used to construct a bias-adjusted within-country estimator for dynamic panels. The simulation evidence reported in Judson and Owen (1999) and Bun and Kiviet (2001) suggests that this estimator performs well relative to standard alternatives when *N* and *T* are small. One minor limitation is that it cannot yet be applied to an unbalanced panel. A more serious limitation, relative to GMM, is that it does not address the possible correlation between the explanatory variables and the disturbances due to simultaneity and measurement error. Nevertheless, for researchers determined to use fixed effects estimation, there is a clear case for implementing this bias adjustment, at least as a complement to other methods.

[60] An alternative approach would be to use small-sample bias adjustments for GMM panel data estimators, such as those described in Hahn, Hausman, and Kuersteiner (2001).

A further issue that arises when estimating dynamic panel data models is that of parameter heterogeneity. If a slope parameter such as β varies across countries, and the explanatory variable is serially correlated, this will induce serial correlation in the error term. If we focus on a simple case where a researcher wrongly assumes $\beta_i = \beta$ for all $i = 1, \ldots, N$ then the error process for a given country will contain a component that resembles $(\beta_i - \beta) \log y_{i,t-1}$. Hence there is serial correlation in the errors, given the persistence of output. The estimates of a dynamic panel data model will be inconsistent even if GMM methods are applied.

This problem was analyzed in more general terms by Robertson and Symons (1992) and Pesaran and Smith (1995) and has been explored in great depth for the growth context by Lee, Pesaran, and Smith (1997,1998). Since an absence of serial correlation in the disturbances is usually a critical assumption for the GMM approach, parameter heterogeneity can be a serious concern. Some of the possible solutions, such as regressions applied to single time series, or the pooled mean group estimator developed by Pesaran, Shin, and Smith (1999), have limitations in studying growth for reasons already discussed. An alternative solution is to split the sample into groups that are more likely to share similar parameter values. Groupings by regional location or level of development are a natural starting point.

Perhaps the state of the art in analyzing growth using panel data and allowing for parameter heterogeneity is represented by Phillips and Sul (2003). They allow for heterogeneity in parameters not only across countries, but also over time. Temporal heterogeneity is rarely investigated in panel studies, but may be important, especially if observed growth patterns combine transitional dynamics towards a country's steady state with fluctuations around that steady-state. Phillips and Sul find some evidence of convergence towards steady states for OECD economies as well as US regions.

We close our discussion of panel data approaches by noting some unresolved issues in their application. It is important to be aware how panel data methods change the substantive interpretation of regression results, and care is needed when moving between the general forms of the estimators and the economic hypotheses under study. Relevant examples occur in analyses of β-convergence. If one finds β-convergence in a panel

study having allowed for fixed effects, the interpretation of this finding is very different than if one finds evidence of convergence in the absence of fixed effects. Specifically, the presence of fixed effects represents an immediate violation of our convergence definitions (20) or (22) as different economies must exhibit steady-state differences in per capita income regardless of whether they have identical saving rates and population growth rates.[61] Fixed effects may even control for the presence of unmodelled determinants of steady state growth, an identification problem analogous to the one that was previously discussed in the context of interpreting the control variables Z in equations (17) and (18) above. Similarly, allowing for differences in time trends for per capita output, as done in Lee, Pesaran, and Smith (1997,1998) means that the finding of extremely rapid β-convergence is consistent with long-run divergence of per capita output across the economies they study; the long-run balanced growth paths are no longer parallel. In an interesting exchange, Lee, Pesaran, and Smith (1998) criticize Islam (1995) for failing to allow for different time trends across countries. In response, Islam (1998) argues that Lee, Pesaran, and Smith are assessing an economically uninteresting form of convergence when they allow for trend differences. This debate is an excellent example of the issues of interpretation that are raised in moving between specific economic hypotheses and more general statistical models.

One drawback of many current panel studies is that the construction of the time series observations can appear arbitrary. There is no inherent reason why 5 or 10 years represent natural spans over which to average observations. Similarly, there is arbitrariness with respect to which time periods are aggregated. A useful endeavor would be the development of tools to ensure that panel findings are robust with respect to the assumptions employed in creating the panel from the raw data.

More fundamentally, the empirical growth literature has not fully addressed the question of the appropriate time horizons over which growth models should be assessed. For example, it remains unclear when business cycle considerations (or instances of output collapses) may be safely ignored when modeling the growth process. While cross-section studies that examine growth over 30-40 year periods might be exempt from this

[61] The impact of controlling fixed effects for interpreting β-convergence is recognized in the conclusion to Islam (1995).

consideration, it is less clear that panel studies employing 5-year averages are genuinely informative about medium-run growth dynamics.

iii. event study approaches

Although we have focused on the limitations of panel data methods, it is clear that the prospects for informative work of this kind should improve over time. The addition of further time periods is valuable in itself, and the history of developing countries in the 1980s and 1990s offers various events that introduce richer time series variation into the data. These events include waves of democratization, macroeconomic stabilization, financial liberalization, and trade liberalization, and panel data methods can be used to investigate their unfolding consequences for growth.

An alternative approach has become popular, and proceeds in a similar way to event studies in the empirical finance literature. In event studies, researchers look for systematic changes in asset returns after a discrete event, such as a profits warning. In fields outside finance, before-and-after studies like this have proved an informative way to gauge the effects of devaluations (see Pritchett (2000a) for references), of inflation stabilization (Easterly (1996)) and the consequences of the debt crisis for investment, as in Warner (1992).

Pritchett (2000a) argues that there is a great deal of scope for studying the growth impact of major events and policy changes in a similar way. The obvious approach is to study the time paths of variables such as output growth, investment and TFP growth, examined before and after such events. In empirical growth research, Henry (2000,2003) has applied this form of analysis to the effects of stock market liberalization on investment and growth, Giavazzi and Tabellini (2004) have considered economic and political liberalizations, while Wacziarg and Welch (2003) have studied the effects of trade liberalization. Depending on the context, one can also study the response of other variables in a way that is informative about the channels of influence. For example, in the case of trade liberalization, it is natural to study the response of the trade share, as in the work of Wacziarg and Welch.

The rigor of this method should not be overplayed. As with any other approach to empirical growth, one has to be cautious about inferring a causal effect. This is clear from exploring the analogy with treatment effects, a focus of recent research in microeconometrics and labor economics.[62] In the study of growth, the treatments – such as democratization – are clearly not exogenously assigned, but are events that have arisen endogenously. Moreover, the treatment effects will be heterogeneous and could depend, for example, on whether a policy change is seen as temporary or permanent (Pritchett (2000a)). In these circumstances, the ability to quantify even an average treatment effect is strongly circumscribed. It may be possible to identify the direction of effects, and here the limited number of observations does have one advantage. With a small number of cases to examine, it is easy for the researcher to present a graphical analysis that allows readers to gauge the extent of heterogeneity in responses, and the overall pattern. At the very least, this offers a useful complement to regression-based methods.

iv. endogeneity and instrumental variables

A final set of data-based issues concerns the identification of instrumental variables in cross-section and time series contexts. An obvious and frequent criticism of growth regressions is that they do little to establish directions of causation. At one level, there is the standard problem that two variables may be correlated but jointly determined by a third. It is very easy to construct growth examples. Variables such as growth and political stability could be seen as jointly determined equilibrium outcomes associated with, say, a particular set of institutions. In this light, a correlation between growth and political stability, even if robust in statistical terms, does not appear especially informative about the structural determinants of growth.

There are many instances in growth research when explanatory variables are clearly endogenously determined (in the economic, not the statistical sense). The most familiar example would be a regression that relates growth to the ratio of investment to

[62]This connection with the treatment effect literature is sometimes explicitly made, as in Giavazzi and Tabellini (2004) and Persson and Tabellini (2003). The connection helps to understand the limitations of the evidence, but the scope for resolving the associated identification problems may be limited in cross-country data sets.

GDP. This may tell us that the investment share and growth are associated, but stops short of identifying a causal effect. Even if we are confident that a change in investment would affect growth, in a sense this just pushes the relevant question further back, to an understanding of what determines investment.

When variables are endogenously determined in the economic sense, there is also a strong chance that they will be endogenous in the technical sense, namely correlated with the disturbances in the structural equation for growth. To give an example, consider what happens if political instability lowers growth, but slower economic growth feeds back into political instability. The estimated regression coefficient will tend to conflate these two effects and will be an inconsistent estimate of the causal effect of instability.[63]

Views on the importance of these considerations differ greatly. One position is that the whole growth research project effectively capsizes before it has even begun, but Mankiw (1995) and Wacziarg (2002) have suggested an alternative view. According to them, one should accept that reliable causal statements are almost impossible to make, but use the partial correlations of the growth literature to rule out some possible hypotheses about the world. Wacziarg uses the example of the negative partial correlation between corruption and growth found by Mauro (1995). Even if shown to be robust, this correlation does not establish that somehow reducing corruption will be followed by higher growth rates. But it does make it harder to believe some of the earlier suggestions, rarely based on evidence, that corruption could be actively beneficial.

One approach is to model as many as possible of the variables that are endogenously determined. A leading example is Tavares and Wacziarg (2001), who estimate structural equations for various channels through which democracy could influence development. In their analysis, democracy affects growth via factors such as its effect on human capital accumulation, physical capital accumulation, inequality and government expenditures. They conclude the net effect of democracy on growth is

[63]Although this 'reverse causality' interpretation of endogeneity is popular and important, it should be remembered that a correlation between an explanatory variable and the error term can arise for other reasons, including omitted variables and measurement error. As we discuss, it is important to bear this more general interpretation of the error term in mind when judging the plausibility of exclusion restrictions in instrumental variable procedures.

slightly negative, despite the positive contributions that are made from the role of democracy in promoting greater human capital and reduced inequality.

This approach has some important advantages in both economic and statistical terms. It can be informative about underlying mechanisms in a way that much empirical growth research is not. From a purely statistical perspective, if the structural equations are estimated jointly by methods such as three stage least squares or full information maximum likelihood, this is likely to bring efficiency gains. That said, systems estimation is not necessarily the best route: it has the important disadvantage that specification errors in one of the structural equations could contaminate the estimates obtained for the others.

The most common response to the endogeneity of growth determinants has been the application of instrumental variable procedures to a single structural equation, with growth as the dependent variable. As mentioned in Section IV, two growth studies that employ instrumental variables estimators based on lagged explanatory variables are Barro and Lee (1994) and Caselli, Esquivel, and Lefort (1996). Appendices 3 and 4 describe a wide range of other instrumental variables that have been proposed for the Solow variables and other growth determinants respectively, where the focus has been on the endogeneity of particular variables. The variety of instruments that have been proposed illustrates that it is relatively straightforward to find an instrument that is correlated with the endogenous explanatory variable(s).

This apparent success may be illusory. In our view, the belief that it is easy to identify valid instrumental variables in the growth context is deeply mistaken. We regard many applications of instrumental variable procedures in the empirical growth literature to be undermined by the failure to address properly the question of whether these instruments are valid, i.e. whether they may be plausibly argued to be uncorrelated with the error term in a growth regression. When the instrument is invalid, instrumental variables estimates will of course be inconsistent. Not enough is currently known about the consequences of "small" departures from validity, but it is certainly possible to envisage circumstances under which ordinary least squares would be preferable to instrumental variables on, say, a mean square error criterion.

A common misunderstanding, perhaps based on confusing the economic and statistical versions of "exogeneity", is that predetermined variables, such as geographical characteristics, are inevitably strong candidates for instruments. There is, however, nothing in the predetermined nature of these variables to ensure either that they are not direct growth determinants or that they are uncorrelated with omitted growth determinants. Even if we take the extreme (from the perspective of being predetermined) example of geographic characteristics, there are many channels through which these could affect growth, and therefore many ways in which they could be correlated with the disturbances in a growth model. Brock and Durlauf (2001a) use this type of reasoning to make a very general critique of the use of instrumental variables in growth economics, basing it on the notion of theory open-endedness that we have described earlier. Since growth theories are mutually compatible, the validity of an instrument requires a positive argument that it cannot be a direct growth determinant or correlated with an omitted growth determinant. For many of the instrumental variables that have been proposed, this is clearly not the case.

Discussions of the validity of instruments inevitably suffer from some degree of imprecision because of the need to make qualitative and subjective judgments. When one researcher claims that it is implausible that a given instrument is valid, unless this claim is made on the basis of a joint model of the instruments and the variable of original interest, another researcher can always simply reject the assertion as unpersuasive. To be clear, this element of subjectivity does not mean that arguments about validity are pointless.[64] Rather, one must recognize that not all statistical questions can be adjudicated on the basis of mathematical analysis.

To see how different instruments might be assigned different levels of plausibility, we consider two examples. Brock and Durlauf (2001a) single out Frankel and Romer's (1999) geographic instruments as an example where instrument validity

[64]Put differently, one does not require a precise definition of what makes an instrument valid in order to argue whether a given instrument is valid or not. To take an example due to Taylor (1998), the absence of a precise definition of money does not weaken my belief that the currency in my wallet is a form of money, whereas the computer on which this paper is written is not. To claim such arguments cannot be made is known as the Socratic fallacy.

appears suspect as such variables are likely correlated with features of a country's economic, political, legal, and social institutions.[65] In our view, the large body of theoretical and empirical evidence on the role of institutions on growth, as well as even a cursory reading of history, renders the orthogonality assumptions required to use the instruments questionable.[66] For example, it is a standard historical claim that the fact that Great Britain is an island had important implications for its political development. While Frankel (2003) suggests that this worry is contrived, the argument against instrument validity flows quite naturally from modern growth theory and the many possible ways in which geographic characteristics such as remoteness could influence development.

As an example where instrument validity may be more plausible, consider Cook (2002a). He employs measures of damage caused by World War II as instruments for various growth regressors such as savings rates. The validity of Cook's instruments again relies on the orthogonality of World War II damage with omitted postwar growth determinants. It may be that levels of wartime damage had consequences for post-War growth performance in other respects (such as institutional change) but this argument is perhaps less straightforward than in the case of geographic characteristics.

To be clear, this discussion is nowhere near sufficient to conclude that Frankel and Romer's instruments are invalid whereas Cook's are valid. Rather our point is that conclusions concerning the relative plausibility of one set of instruments versus another need to rest on explicit arguments. It is not enough to appeal to a variable being predetermined, because this does not ensure that it is uncorrelated with the disturbances in the structural equation being estimated. A key implication of our discussion is that historical information has a vital role to play in facilitating formal growth analyses and evaluating exclusion restrictions.

---

[65]While questions about the validity of instrumental variables arise in virtually all contexts, the force of these concerns differs across contexts. For example, in rational expectations models, lagged variables are natural instruments with respect to variables that, from the perspective of the theoretical model, are martingale differences, as occurs for excess holding returns. Objections to particular instruments in these contexts typically rely on alternative specifications of preferences or some other modification of the economic logic of the original model. This is quite different from the openendedness of growth theories.

[66]The body of work on institutions and growth excellently summarized in Acemoglu, Johnson, and Robinson (2004) is strongly supportive of this claim.

This discussion of instrumental variables indicates another important, albeit neglected, issue in empirical growth analysis: the relationship between model specification and instrumental variable selection. One cannot discuss the validity of particular instruments independently from the choice of the specific growth determinants under study. An important outstanding research question is whether model uncertainty and instrumental variable selection can be integrated simultaneously into some of the methods we have described, including model averaging and automated model selection. The recent work of Hendry and Krolzig (2005) on automated methods includes an ambitious approach to systematic model selection for simultaneous equation models in which identifying restrictions are determined by the data.

VII. Econometric issues II: data and error properties

In this section we consider a range of questions that arise in growth econometrics from the properties of data and errors. Starting with data issues, Section VII.i examines how one may handle outliers in growth data. Section VII.ii examines the problem of measurement error. This is an important issue since there are good reasons to believe that the quality of the data is sometimes poor for less developed economies. In Section VII.iii we consider the case where data are not even measured, i.e are missing. Turning to issues of the properties of model errors, Section VII.iv examines the analysis of heteroskedasticity in growth contexts. Finally, Section VII.v addresses the problem of cross-section correlation in model errors.

i. outliers

Empirical growth researchers often work with small data sets and estimate relatively simple models. In these circumstances, OLS regressions are almost meaningless unless they have been accompanied by systematic investigation of the data, including the sensitivity of the results to outlying observations.

There are various reasons why some observations may be unrepresentative. It is possible for variables to be measured with error for that particular region or country. Alternatively, the model specified by the researcher may omit a relevant consideration, and so a group of country observations will act as outliers. By construction, least squares estimates can be highly sensitive to the presence of small groups of observations. The practical implication is that OLS can give a misleading account of the patterns in the majority of the data. The dangers of using OLS were forcibly expressed by Swartz and Welsch (1986, p. 171): "In a world of fat-tailed or asymmetric error distributions, data errors, and imperfectly specified models, it is just those data in which we have the least faith that often exert the most influence on the OLS estimates".

Some researchers respond to this concern using leverage measures or single-case diagnostics such as Cook's distance statistic. There are well-known problems with these approaches, because where more than one outlier is present, its effect can easily be hidden by another (known in the statistics literature as "masking"). By far the best response is to use a more robust estimator, such as least trimmed squares, at least as a preliminary way of investigating the data.[67] These issues are discussed in more detail in Temple (1998,2000b).

ii. measurement error

We now turn to a more general discussion of measurement error. It is clear that measurement errors are likely to be pervasive, especially in data that relate to developing countries. Concepts that appear straightforward in economic models can present huge measurement problems in practice, as in the example of the capital stock discussed by Pritchett (2000b). Yet relatively few empirical studies of growth consider the impact of measurement error in any detail.

[67]This estimator should not be confused with trimmed least squares, and other methods based on deleting observations with high residuals in the OLS estimates. A residual-based approach is inadequate for obvious reasons.

The best-known statistical result applies to a bivariate model where the independent variable is measured with error.[68] The estimate of the slope coefficient will be biased towards zero, even in large samples, because measurement error induces covariance between the observable form of the regressor and the error term. This attenuation bias is well known, but sometimes misleads researchers into suggesting that measurement error will only mask effects, a claim that is not true in general. When there are multiple explanatory variables, but only one is measured with error, then typically all the parameter estimates will be biased. Some parameter estimates may be biased away from zero and, although the direction of the bias can be estimated consistently, this is rarely done. When several variables are measured with error, the assumption that measurement error only hides effects is even less defensible.

Where measurement error is present, the coefficients are typically not identified unless other information is used. The most popular solution is to use instrumental variables, if an instrument can be found which is likely to be independent of the measurement error. A more complex solution is to exploit higher-order sample moments to construct more sophisticated estimators, as in Dagenais and Dagenais (1997). These procedures may be unreliable in small samples since the use of higher-order moments will make them especially sensitive to outliers.

Sometimes partial identification is possible, in the sense that bounds on the extent of measurement error can be used to derive consistent estimates of bounds on the slope parameters. Although it can be difficult for researchers to agree on sensible bounds on the measurement error variances, there are easier ways of formulating the necessary restrictions, as discussed by Klepper and Leamer (1984). Their reverse regression approach was implemented by Persson and Tabellini (1994) and Temple (1998), but has rarely been used by other researchers. Another strategy is to investigate sensitivity to varying degrees of measurement error, based on method-of-moments corrections. Again, this is easy to implement in linear models, and should be applied more routinely than it is at present. Temple (1998) provides a discussion of both approaches in the context of

[68]This and the following discussion assume classical measurement error. Under more general assumptions, it is usually even harder to identify the consequences of measurement error for parameter estimates.

estimating technology parameters and the rate of conditional convergence within the Mankiw, Romer, and Weil (1992) model.

iii. missing data

Some countries never appear in growth data sets, partly by design: it is common to leave out countries with very small populations, oil producers, and transition economies. These are countries that seem especially unlikely to lie on a regression surface common to the majority of the OECD countries or the developing world. Countries with small populations should not be allowed to carry a great deal of weight in attempting to draw generalizations about growth for larger countries.

Other countries are left out for different reasons. When a nation experiences political chaos, or lacks economic resources, the collection of national accounts statistics will be a low priority. This means that countries like Afghanistan, Ethiopia and Somalia rarely appear in comparative growth studies. In other cases, countries appear in some studies but not in others, depending on the availability of particular variables of interest.

Missing data are of course a potentially serious problem. If one started from a representative data set and then deleted countries at random, this would typically increase the standard errors but not lead to biased estimates. More serious difficulties arise if countries are missing in a systematic way, because then parameter estimates are likely to be biased. This problem is given relatively little attention in mainstream econometrics textbooks, despite a large body of research in the statistics literature.

A variety of solutions are possible, with the simplest being one form or another of imputation, with an appropriate adjustment to the standard errors. Hall and Jones (1999) and Hoover and Perez (2004) are among the few empirical growth studies to carry out imputation in a careful and systematic way. This approach may be especially useful when countries are missing from a data set because a few variables are not observed for their particular cases. It is then easy to justify using other available information to predict the missing data, and thereby exploit the additional information in the variables that are observed. Alternative approaches to missing data are also available, based on likelihood or Bayesian methods, which can be extended to handle missing observations.

iv. heteroskedasticity

It is common in cross-section regressions for the underlying disturbances to have a non-constant variance. As is well known, the coefficient estimates remain unbiased, but OLS is inefficient and the estimates of the standard errors are biased. Most empirical growth research simply uses the heteroskedasticity-consistent standard errors developed by Eicker (1967) and White (1980). These estimates of the standard errors are consistent but not unbiased, which suggests that alternative solutions to the problem may be desirable. For data sets of the size found in cross-country empirical work, the alternative estimators developed by MacKinnon and White (1985) are likely to have better finite sample properties, as discussed in Davidson and MacKinnon (1993) and supported by simulations in Long and Ervin (2000).

There are at least two other concerns with the routine application of White's heteroskedasticity correction as the only response to heteroskedasticity. The first is that by exploiting any structure in the variance of the disturbances, using weighted least squares, it may be possible to obtain efficiency gains. The second and more fundamental objection is that heteroskedasticity can often arise from serious model misspecification, such as omitted variables or neglected parameter heterogeneity. Evidence of heteroskedasticity should then prompt revisions of the model for the conditional mean, rather than mechanical adjustments to the standard errors. See Zietz (2001) for further discussion and references.

v. cross-section error correlation

An unresolved issue in growth econometrics is the treatment of cross-section correlation in model errors. Such correlation may have important consequences for inference; as noted by DeLong and Summers (1991) in the growth context, failure to account for cross-sectional dependence can lead to incorrect calculation of standard errors and hence, incorrect inferences. One would certainly expect cross-sectional dependence to be present when studying growth. For example, countries that are geographically close together, or trading partners, may experience common shocks.

Whether this effect is sizeable remains an open question, but one that might be addressed using ideas developed in Baltagi et al. (2003) and Driscoll and Kraay (1998), among others. In the context of growth regressions, work on cross-section dependence may be divided into two lines. One direction concerns the identification of the presence of cross-section dependence. Pesaran (2004a) develops tests for cross-section dependence that do not rely on any prior ordering; this framework in essence sums the cross-section sample error correlations in a panel and evaluates whether they are consistent with the null hypothesis that the population correlations are zero. Specifically, he proposes (recalling that N denotes the cross-section dimension and T the time dimension) a cross-section dependence statistic CD

$$CD = \sqrt{\frac{2T}{N(N-1)}}\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\hat{\rho}_{i,j}\right) \tag{66}$$

where $\hat{\rho}_{i,j}$ is the sample correlation between $\varepsilon_{i,t}$ and $\varepsilon_{j,t}$; Pesaran gives conditions under which this statistic converges to a Normal (0,1) random variable (as N and T become infinite) under the null hypothesis of no cross-section correlation. This test statistic is based on earlier work by Breusch and Pagan (1980) and appears to possess good finite sample properties in comparison to this earlier work. Using a country-level panel, Pesaran (2004a) finds strong rejections of the null of no cross-section correlation both for the world as a whole as well as within several geographic groupings.

The second and primary direction for the analysis of cross-section correlation has been concerned not so much with testing for its presence, but rather accounting for its presence in growth exercises. One approach relies on formulating a statistical model of the dependence. Phillips and Sul (2003) model the residuals in a growth panel as

$$\varepsilon_{i,t} = \delta_i \theta_t + u_{i,t} \tag{67}$$

where θ_t and $u_{i,t}$ are independent random variables; $u_{i,t}$ is assumed to be i.i.d. across countries and across time. Phillips and Sul (2002) describe the properties of panel estimators under this assumption.

Another possibility in analyzing cross-sectional dependence is to treat the problem as one of spatial correlation in errors. The problem of spatial correlation has been much studied in the regional science literature, and statisticians in this field have developed spatial analogues of many time series concepts, see Anselin (2001) for an overview. Spatial methods have, in our view, an important role to play in growth econometrics. However, when these methods are adapted from the spatial statistics literature, they raise the problem of identifying the appropriate notion of space. One can imagine many reasons for cross-section correlation. If one is interested in technological spillovers, it may well be the case that in the space of technological proximity, the United Kingdom is closer to the United States than is Mexico. Put differently, unlike the time series and spatial cases, there is no natural cross-section ordering to elements in the standard growth data sets. Following language due to Akerlof (1997) countries are perhaps best thought of as occupying some general socio-economic-political space defined by a range of factors; if one could identify their locations, then spatial methods could be implemented.

An interesting approach to addressing the relevant spatial location of countries is pursued by Conley and Ligon (2002). In their analysis, they attempt to construct estimates of the spatial covariation of the residuals ε_i in a cross-section. In order to do this, they construct different measures of socioeconomic distance between countries. They separately consider geographic distance (measured between capital cities), as well as measures of the costs of transportation between these cities. Once a distance metric is constructed, these are used to construct a residual covariance matrix. Estimation methods for this procedure are developed in Conley (1999). Conley and Ligon (2002) find that allowing for cross-section dependence in this way is relatively unimportant in terms of appropriate calculation of standard errors for growth model parameters. Their methods could be extended to allow for comparisons of different variables as the source for cross-section correlation as is done in Conley and Topa (2002) in the context of residential neighborhoods. A valuable generalization of this work would be the modeling of cross-

section correlations as a function of multiple variables. Such an analysis would make further progress on the measurement of distances in socioeconomic space, which, as we have suggested, presumably are determined by multiple channels.

A generally unexplored possibility for studying cross-section dependence in growth (and other contexts) is to model these correlations structurally as the outcome of spillover effects.[69] The theoretical literature on social interactions studies cross-sectional dependences in precisely this way (see Brock and Durlauf (2001b) for a survey of this literature). While such models have the potential for providing firm microfoundations for cross-section dependence, the presence of such spillovers has consequences for identification that are not easily resolved (Brock and Durlauf (2001b), Manski (1993)) and which have yet to be explored in growth contexts; Binder and Pesaran (2001) and Brock and Durlauf (2001b) analyze identification and estimation problems for intertemporal environments that are particularly germane to growth contexts.

VIII. Conclusions: the future of growth econometrics

In this section, we offer some closing thoughts on the most promising directions for empirical growth research. We are not the first authors to set out manifestoes for the field, and we explicitly draw on previous contributions, many of which deserve wider currency. It is also interesting to compare the current state of the field against the verdicts offered in the early survey by Levine and Renelt (1991). One dominant theme will be that the empirical study of growth requires an eclectic approach, and that the field has been harmed by a tendency for research areas to evolve independently, without enough interaction.[70] This is not simply a question of using a variety of techniques: it also means that there needs to be a closer connection between theory and evidence, a willingness to

[69] An exception is Easterly and Levine (1997b).

[70] To give a specific example, the macroeconomic literature on international technology differences only rarely acknowledges relevant work by trade economists, including estimates of the Heckscher-Ohlin-Vanek model that suggest an important role for technology differences. See Klenow and Rodriguez-Clare (1997b) for more discussion.

draw on ideas from areas such as trade theory, and more attention to particular features of the countries under study.

We start with Pritchett (2000a), who lists three questions for growth researchers to address:

- What are the conditions that initiate an acceleration of growth or the conditions that set off sustained decline?
- What happens to growth when policies – trade, macroeconomic, investment – or politics change dramatically in episodes of reform?
- Why have some countries absorbed and overcome shocks with little impact on growth, while others seem to have been overwhelmed by adverse shocks?

This agenda seems to us very appropriate, not least because it focuses attention on substantive economic issues rather than the finer points of estimating aggregate production functions. The importance of the first of Pritchett's questions is evident from the many instances where countries have moved from stagnation to growth and vice versa. A paper by Hausmann, Pritchett, and Rodrik (2004) explicitly models transitions to fast growth ("accelerations") and makes clear the scope for informative work of this kind. The second question we have discussed above, and research in this vein is becoming prominent, as in Henry (2000, 2003), Giavazzi and Tabellini (2004), and Wacziarg and Welch (2003). Here, one of the major challenges will be to relax the (sometimes only implicit) assumption that policies are randomly assigned. Finally, an important paper by Rodrik (1999) has addressed the third question, namely what determines varying responses to major shocks.

In all three cases, it is clear that econometric work should be informed by detailed studies of individual countries, such as those collected in Rodrik (2003). Too much empirical growth research proceeds without enough attention to the historical and institutional context. For example, a newcomer to this literature might be surprised at the paucity of work that integrates growth regression findings with, say, the known consequences of the 1980s debt crisis.

Another reason for advocating case studies is that much of the empirical growth literature essentially points only to reduced-form partial correlations. These can be useful, but it is clear that we often need to move beyond this. A partial correlation is more persuasive if it can be supported by theoretical arguments. The two combined are more persuasive if there is evidence of the intermediating effects or mechanisms that are emphasized in the relevant theory. There is plenty of scope for informative work that tries to isolate mechanisms by which variables such as financial depth, inequality, and political institutions shape the growth process. Wacziarg (2002), in particular, highlights the need for a structural growth econometrics, one that aims to recover channels of causation, and hence supports (or undermines) the economic significance of the partial correlations identified in the literature.

A more extreme view is that growth econometrics should be supplanted by the calibration of theoretical models. Klenow and Rodriguez-Clare (1997b) emphasize the potential of such an approach and note that Mankiw, Romer, and Weil's (1992) influential analysis can be seen partly as a comparison of estimated parameter values with those associated with specific theoretical models. Relatively little of the empirical work that has followed has achieved a similarly close connection between theory and evidence, and this has been a recurring criticism of the literature (for example, Levine and Renelt (1991) and Durlauf (2001)).

It may be premature to say that econometric approaches should be entirely replaced by calibration exercises, but the two methods could surely inform each other more often than at present. Calibrated models can help to interpret parameter estimates, not least in comparing the magnitude of the estimates with the implications of plausible models. Klenow and Rodriguez-Clare (1997b) discuss examples of this in more detail. At the same time, the partial correlations identified in growth econometrics can help to act as a discipline on model-building and can indicate where model-based quantitative investigations are most likely to be fruitful. This role for growth econometrics is likely to be especially useful in areas where the microeconomic evidence used to calibrate structural models is relatively weak, or the standard behavioral assumptions may be flawed.

The need for a tighter connection between theory and evidence is especially apparent in certain areas. The workhorse model for many empirical growth papers continues to be Solow-Swan, a closed economy model which leaves out aspects of interdependence that are surely important. Howitt (2000) has shown that growth regression evidence can be usefully reinterpreted in the light of a multi-country theoretical model with a role for technology diffusion. More generally, there is a need for researchers to develop empirical growth frameworks that acknowledge openness to flows of goods, capital and knowledge. These issues are partly addressed by the theoretical analysis of Barro, Mankiw, and Sala-i-Martin (1995) and empirical work that builds on such ideas deserves greater prominence. Here especially, research that draws on the quantitative implications of specific models, as in the work of Eaton and Kortum (1999, 2001) on technology diffusion and the role of imported capital goods, appears to be an important advance.

The neglect of open economy aspects of the countries under study is mirrored elsewhere. Much of the empirical literature uses a theoretical framework that was originally developed to explain the growth experiences of the USA and other developed nations. Yet this framework is routinely applied to study developing countries, and there appears plenty of scope for models that incorporate more of the distinctive features of poorer countries. These could include the potentially important roles of agricultural employment, dualism, and structural change, and in some cases, extensive state involvement in production. This is an area in which empirical growth researchers have really only scratched the surface.

Some of these issues are connected to an important current research agenda, namely the need to distinguish between different types of growth and their distributional consequences. For example, the general equilibrium effects of productivity improvements in agriculture may be very different to those in services and industry. Identifying the nature of "pro-poor" growth will require more detailed attention to particular features of developing countries. Given that the main source of income for the poor is usually labour income, growth researchers will need to integrate their models with theory and evidence from labour economics, in order to study how growth and labour markets interact.

Agénor (2004) considers some of the relevant issues, and again this appears to be a vital direction for future research.

Ideally, research along these various lines will utilize not only statistics, but also the power of case studies in generating hypotheses, and in deepening our understanding of the economic, social and political forces at work in determining growth outcomes. Case studies may be especially valuable in two areas. The first of these is the study of technology transfer. As emphasized in the survey by Klenow and Rodriguez-Clare (1997b), we do not know enough about why some countries are more successful than others in climbing the "ladder" of product quality and technological complexity. What are the relative contributions of human capital, foreign direct investment and trade? In recent years some of these issues have been intensively studied at the microeconomic level, especially the role of foreign direct investment and trade, but there remains work to be done in mapping firm and sector-level evidence into a set of aggregate implications.

A second area in which case studies are likely to prove valuable is the study of political economy, in its modern sense. It is a truism that economists, particularly those considering development, have become more aware of the need to account for the two-way interaction between economics and politics. A case can be made that the theoretical literature has outpaced the empirical literature in this regard. Studies of individual countries, drawing on both economic theory and political science, would help to close this gap.

Thus far, we have highlighted a number of limitations of existing work, and directions in which further research seems especially valuable. Some of the issues we have considered were highlighted much earlier by Levine and Renelt (1991), and that might lead to pessimism over the long-term prospects of this literature.[71] This also shows that our prescriptions for future research could seem rather pious, since the improvements we recommend are easier said than done. We end our review by considering some areas in which genuine progress has been made, and where further progress appears likely.

One reason for optimism is the potential that recently developed model averaging

[71]Only now are researchers beginning to engage with some of the issues they raised, such as the varying conditions under which it is appropriate to use international rather than national prices in making productivity comparisons and constructing capital stocks.

methods have for shedding new light on growth questions. These methods help to address the model selection and robustness issues that have long been identified as a major weakness of cross-country growth research. By framing the problem explicitly in terms of model uncertainty, in the way envisaged by Leamer (1978), it is possible to consider many candidate explanatory variables simultaneously, and identify which effects appear to be systematic features of the data, as reflected in posterior probabilities of inclusion. The Bayesian approach to model averaging also provides an index of model adequacy, the posterior model probability, that is easy to interpret, and that allows researchers to gauge the extent of overall model uncertainty. Above all, researchers can communicate the degree of support for a particular hypothesis with more faith that the results do not depend on an arbitrary choice of regression specification. Although the application of Bayesian model averaging inevitably has limitations of its own, it appears more rigorous than many of the alternatives, and we expect a number of familiar growth questions to be revisited using these methods.

Another reason for optimism is that the quality of available data is likely to improve over time. The development of new and better data has clearly been one of the main achievements of the empirical growth literature since the early 1990s, and one that was not foreseen by critics of the field. Researchers have developed increasingly sophisticated proxies for drivers of growth that appeared resistant to statistical analysis. One approach, pioneered in the growth literature by Knack and Keefer (1995) and Mauro (1995), has been country-specific ratings compiled by international agencies. Such data increasingly form the basis for measures of corruption, government efficiency, and protection of property rights. More recent work such as that of Kaufmann et al. (1999a,1999b,2003) has established unusually comprehensive measures of various aspects of institutions.

The construction of proxies is likely to make increasing use of latent variable methods. These aim to reduce a set of observed variables to a smaller number of indicators that are seen as driving the majority of the variation in the original data, and that could represent some underlying variable of interest. For example, the extent of democracy is not directly observed, but is often obtained by applying factor analysis or extracting principal components from various dimensions of political freedom. There are

obvious dangers with this approach, but the results can be effective proxies for concepts that are otherwise hard to measure.[72] They also help to overcome the dimensionality problem associated with cross-country empirical work. To be successfully employed, the rigorous use of a latent variable as a regressor will generally need to acknowledge the presence of measurement error.[73]

Using latent variables makes especially good sense under one view of the proper aims of growth research. It is possible to argue that empirical growth studies will never give good answers to precise hypotheses, but can be informative at a broader level. For example, a growth regression is unlikely to tell us whether the growth effect of inflation is more important than the effect of inflation uncertainty, because these two variables are usually highly correlated. It may even be difficult to distinguish the effects of inflation from the effects of sizeable budget deficits.[74] Instead a growth regression might be used to address a less precise hypothesis, such as the growth dividend of macroeconomic stability, broadly conceived. In this context, it is natural to use latent variable approaches to measure the broader concept.

Another valuable development is likely to be the creation of rich panel data sets at the level of regions within countries. Regional data offer greater scope for controlling for some variables that are hard to measure at the country level, such as cultural factors. By comparing experiences across regions, there may also be scope for identifying events that correspond more closely to natural experiments than those found in cross-country data. Work such as that by Besley and Burgess (2000,2002,2004) using panel data on Indian

[72]A relevant question, not often asked, is how high the correlation between the proxy and the true predictor has to be for the estimated regression coefficient on the proxy to be of the "true" sign. Krasker and Pratt (1986,1987) have developed methods that can be used to establish this under surprisingly general assumptions.

[73]In principle this can be addressed by structural equation modeling, using software like EQS or LISREL to estimate a system of equations that includes explicit models for latent variables, an approach used elsewhere in the social sciences. Most economists are not familiar with this approach, and this makes the assumptions and results hard to communicate. It is therefore not clear that a full latent variable model should be preferred to a simpler solution, such as one of those we discuss in the measurement error section above.

[74]As Sala-i-Martin (1991) has argued, various specific indicators of macroeconomic instability should perhaps be seen as symptoms of some deeper, underlying characteristic of a country.

states shows the potential of such an approach. In working with such data more closely, one of the main challenges will be to develop empirical frameworks that incorporate movements of capital and labour between regions: clearly, regions within countries should only rarely be treated as closed economies. Shioji (2001b) is an example of how analysis using regional data can take this into account.

Even with better data, at finer levels of disaggregation, the problem of omitted variables can only be alleviated, not resolved. It is possible to argue that the problem applies equally to historical research and case studies, but at least in these instances, the researcher may have some grasp of important forces that are difficult to quantify. Since growth researchers naturally gravitate towards determinants of growth that can be analyzed statistically, there is an ever-present danger that the empirical literature, even taken as a whole, yields a rather partial and unbalanced picture of the forces that truly matter. Even a growth model with high explanatory power, in a statistical sense, has to be seen as a rather provisional set of ideas about the forces that drive growth and development.

This brings us to our final points. We once again emphasize that empirical progress on the major growth questions requires attention to the evidence found in qualitative sources such as historical narratives and studies by country experts. One example we have given in the text concerns the validity of instrumental variables: understanding the historical experiences of various countries seems critical for determining whether exclusion restrictions are plausible. In this regard work such as that of Acemoglu, Johnson, and Robinson (2001,2002) is exemplary. More generally, nothing in the empirical growth literature suggests that issues of long-term development can be disassociated from the historical and cultural factors that fascinated commentators such as Max Weber. Where researchers have revisited these issues, as in Barro and McCleary (2003), the originality resides less in the conception of growth determinants and more in the scope for new statistical evidence. Of course, the use of historical analysis also leads back to the value of case studies, a point that has recurred throughout this discussion.

In conclusion, growth econometrics is an area of research that is still in its infancy. To its credit, the field has evolved in response to the substantive economic questions that arise in growth contexts. The nature of the field has also led

econometricians to introduce a number of statistical methods into economics, including classification and regression tree algorithms, robust estimation, threshold models and Bayesian model averaging, that appear to have wide utility. As with any new literature, especially one tackling questions as complex as these, it is possible to identify significant limitations of the existing evidence and the tools that are currently applied. But the progress that has been made in growth econometrics in the brief time since its emergence gives reason for continued optimism.

Appendix 1: Data

Key to the 102 countries

AGO, Angola, ARG, Argentina, AUS, Australia, AUT, Austria, BDI, Burundi, BEL, Belgium, BEN, Benin, BFA, Burkina Faso, BGD, Bangladesh, BOL, Bolivia, BRA, Brazil, BWA, Botswana, CAF, Central African Republic, CAN, Canada, CHE, Switzerland, CHL, Chile, CHN, China, CIV, Cote d'Ivoire, CMR, Cameroon, COG, Rep. of Congo, COL, Colombia, CRI, Costa Rica, CYP, Cyprus, DNK, Denmark, DOM, Dominican Republic, ECU, Ecuador, EGY, Egypt, ESP, Spain, ETH, Ethiopia, FIN, Finland, FJI, Fiji, FRA, France, GAB, Gabon, GBR, United Kingdom, GHA, Ghana, GIN, Guinea, GMB, The Gambia, GNB, Guinea-Bissau, GRC, Greece, GTM, Guatemala, GUY, Guyana, HKG, Hong Kong, HND, Honduras, IDN, Indonesia, IND, India, IRL, Ireland, IRN, Iran, ISR, Israel, ITA, Italy, JAM, Jamaica, JOR, Jordan, JPN, Japan, KEN, Kenya, KOR, Rep. of Korea, LKA, Sri Lanka, LSO, Lesotho, MAR, Morocco, MDG, Madagascar, MEX, Mexico, MLI, Mali, MOZ, Mozambique, MRT, Mauritania, MUS, Mauritius, MWI, Malawi, MYS, Malaysia, NAM, Namibia, NER, Niger, NGA, Nigeria, NIC, Nicaragua, NLD, Netherlands, NOR, Norway, NPL, Nepal, NZL, New Zealand, PAK, Pakistan, PAN, Panama, PER, Peru, PHL, Philippines, PNG, Papua New Guinea, PRT, Portugal, PRY, Paraguay , ROM, Romania, RWA, Rwanda, SEN, Senegal, SGP, Singapore, SLV, El Salvador, SWE, Sweden, SYR, Syria, TCD, Chad, TGO, Togo, THA, Thailand, TTO, Trinidad & Tobago, TUR, Turkey, TWN, Taiwan, TZA, Tanzania, UGA, Uganda, URY, Uruguay, USA, USA, VEN, Venezuela, ZAF, South Africa, ZAR, Dem. Rep. Congo, ZMB, Zambia, ZWE, Zimbabwe

Extrapolation

Where data on GDP per worker for the year 2000 are missing from PWT 6.1, but are available for 1996 or after, we extrapolate using the growth rate between 1990 and the latest available year. This procedure helps to alleviate the biases that can occur when countries are missing from the sample for systematic reasons, such as political or economic collapse.

The countries involved are Angola (extrapolated from 1990-1996), Botswana (1999), Central African Republic (1998), Democratic Republic of Congo (1997), Cyprus (1996), Fiji (1999), Guyana (1999), Mauritania (1999), Namibia (1999), Papua New Guinea (1999), Singapore (1996), and Taiwan (1998).

Appendix 2: Variables in Cross-Country Growth Regressions

+/- = sign of coefficient in the corresponding growth regression
? = sign not reported
* = claimed to be significant
_ = claimed to be insignificant

R.H.S. Variables		Studies
Capitalism		• Hall and Jones (1999) (+,*)
Capital account liberalization		• Eichengreen and Leblang (2003) (+,*)
Corruption		• Mauro (1995) (-,*) • Welsch (2003) (-,*)
Democracy	Minimum levels	• Barro (1996) (1997) (+,*)
	...Higher levels	• Barro (1996) (1997) (-,*)
	Overall	• Alesina et al. (1996) (?,_) • Minier (1998) (+,*)
	'Voice'	• Dollar and Kraay (2003) (-,*)
Demographic Characteristics	Share of Population 15 or below	• Barro and Lee (1994) (-,*)
	Share of Population 65 or over	• Barro and Lee (1994) (?,_)
	Growth of 15-65 population share	• Bloom and Sachs (1998) (+,*)
Education	College Level	• Barro and Lee (1994) (-,_)
	Female (level)	• Barro and Lee (1994) (-,*) • Barro (1996) (1997) (-,*) • Caselli, et al. (1996) (+,*) • Forbes (2000) (-,*)
	Female (growth)	• Barro and Lee (1994) (-,*)
	Male (level)	• Barro and Lee (1994) (+,*) • Barro (1996) (+,*) • Caselli, et al. (1996) (-,*) • Forbes (2000) (+,*)
	Male (growth)	• Barro and Lee (1994) (+,*)

	Overall (level)	• Azariadis and Drazen (1990) (+,*) • Barro (1991) (+,*) • Knowles and Owen (1995) (+,_) • Easterly and Levine (1997a) (+,*) • Krueger and Lindahl (2000) (+,*) • Bils and Klenow (2000) (+,*)
	Primary Level	• Sachs and Warner (1995) (+,_) • Barro (1997) (-,_)
	Secondary Level	• Sachs and Warner (1995) (+,_)
	Initial Income * Male Schooling	• Barro (1997) (-,*)
	Proportion of Engineering Students	• Murphy, et al. (1991) (+,*)
	Proportion of Law Students	• Murphy, et al. (1991) (-,*)
Ethnicity and Language	Ethno-Linguistic Fractionalization	• Easterly and Levine (1997a) (-,*) • Sala-i-Martin (1997a,b) (?,_) • Alesina, et al. (2003) (-,*)
	Language Diversity	• Masters and McMillan (2001) (-,*/_)
Fertility		• Barro (1991) (1996) (1997) (-,*) • Barro and Lee (1994) (-,*)
Finance	Stock Markets	• Levine and Zervos (1998) (+,*) • Beckaert, et al. (2001) (+,*) • Beck and Levine (2004) (+,*)
	Banks	• Beck and Levine (2004) (+,*)
	Dollarization	• Edwards and Magendzo (2003) (+,_)
	Depth	• Berthelemy and Varoudakis (1995) (+,*) • Odedokun (1996) (+,*) • Ram (1999) (+,_) • Rousseau and Sylla (2001) (+,*) • Deidda and Fattouh (2002) (+,_)

			• Demetriades and Law (2004) (+,*)
	Competition*development		• Claessens and Laeven (2003) (+,*)
	Repression		• Roubini and Sala-i-Martin (1992) (-,*) • Easterly (1993) (-,*)
	Sophistication		• King and Levine (1993) (+,*) • Levine and Zervos (1993) (+,robust) • Easterly and Levine (1997a) (+,*) • Sala-i-Martin (1997a,b) (?,_)
	Credit	Growth rate	• Levine and Renelt (1992) (+,not robust) • De Gregorio and Guidotti (1995) (+,*)
		Volatility	• Levine and Renelt (1992) (+,not robust)
Foreign Direct Investment			• Blonigen and Wang (2004) (+,_)
Fraction of mining in GDP			• Hall and Jones (1999) (+,*)
Geography	Absolute Latitude		• Sala-i-Martin (1997a,b) (+,*) • Bloom and Sachs (1998) (+,*) • Masters and McMillan (2001) (+,_) • Easterly and Levine (2001) (+,*) • Rodrik et al. (2004) (+,*)
	Disease Ecology		• McCarthy, et al. (2000) (+,*) • McArthur and Sachs (2001) (+,*) • Easterly and Levine (2002) (-,*) • Sachs (2003) (-,*)
	Frost days		• Masters and McMillan (2001)(+,*) • Masters and Sachs (2001) (+,*)

	Land locked	• Easterly and Levine (2001) (-,*)
	Coastline (length)	• Bloom and Sachs (1998) (+,*) • Masters and Sachs (2001) (+,*) • Bloom, et al. (2003) (+,*)
	Arable land	• Masters and Sachs (2001) (+,*)
	Rainfall	• Masters and Sachs (2001) (+,*) • Bloom, et al. (2003) (+,*)
	Variance of Rainfall	• Bloom, et al. (2003) (-,*)
	Maximum Temperature	• Bloom, et al. (2003) (-,*)
Government	Consumption (growth)	• Kormendi and Meguire (1985) (+,_)
	Consumption (level)	• Barro (1991) (-,*) • Sachs and Warner (1995) (-,*) • Barro (1996) (-,*) • Caselli, et al. (1996) (+,*) • Barro (1997) (-,*) • Acemoglu, et al. (2002) (-,_)
	Deficits	• Levine and Renelt (1992) (-,not robust) • Fischer (1993) (-,*) • Nelson and Singh (1994) (+,_) • Easterly and Levine (1997a) (-,*) • Bloom and Sachs (1998) (+,*)
	Investment	• Barro (1991) (+,_) • Sala-i-Martin (1997a,b) (?,_) • Kelly (1997) (+,*)
	Various Expenditures	• Levine and Renelt (1992) (-,not robust)
	Military Expenditures	• Aizenman and Glick (2003) (-,*) • Guaresma and Reitschuler (2003) (-,*)
	Military Expenditures under threat	• Aizenman and Glick (2003) (+,*)
	Various Taxes	• Levine and Renelt (1992) (?,not

		robust)
Growth Rate	of the G-7 Countries	• Alesina, Ozler, Roubini, and Swagel (1996) (+,*)
	in the Previous Period	• Easterly, et al. (1993) (+,_) • Alesina, et al. (1996) (+,*/_)
Health	Life expectancy	• Barro and Lee (1994) (+,*) • Bloom and Malaney (1998) (+,*) • Bloom and Sachs (1998) (+,*) • Bloom and Williamson (1998) (+,*) • Hamoudi and Sachs (1999) (+,*) • Gallup et al. (2000) (+,*)
	Change in Malaria Infection Rate	• Gallup, Mellinger and Sachs (2000).
	Adult Survival Rate	• Bhargava et al. (2001)
Industrial Structure	% Small and Medium Enterprises	• Beck, et al. (2003) (+,_)
	Ease of entry and exit	• Beck, et al. (2003) (+,*)
Inequality	Democratic Countries	• Persson and Tabellini (1994) (-,*)
	Non-Democratic Countries	• Persson and Tabellini (1994) (+,_)
	Overall	• Alesina and Rodrik (1994) (-,*) • Forbes (2000) (+,*) • Knowles (2001) (-,*)
Inflation	Growth	• Kormendi and Meguire (1985) (-,*)
	Level	• Levine and Renelt (1992) (-,not robust) • Levine and Zervos (1993) (?,not robust) • Barro (1997) (-,*) (in the range above 15%) • Bruno and Easterly (1998) (-,*) • Motley (1998) (-,*)

		• Li and Zou (2002) (-,*)
	Variability	• Levine and Renelt (1992) (-,not robust) • Fischer (1993) (-,*) • Barro (1997) (+,_) • Sala-i-Martin (1997a,b) (?,_)
Infrastructure Proxies		• Hulten (1996) (+,*) • Easterly and Levine (1997a) (+,*) • Esfahani and Ramirez (2003) (+,*)
Initial Income		• Kormendi and Meguire (1985) (-,*) • Barro (1991) (-,*) • Sachs and Warner (1995) (-,*) • Harrison (1996) (?,_) • Barro (1997) (-,*) • Easterly and Levine (1997a)
Investment Ratio		• Barro (1991) (+,*) • Barro and Lee (1994) (+,*) • Sachs and Warner (1995) (+,*) • Barro (1996) (+,_) • Caselli, et al. (1996) (+,*) • Barro (1997) (+,_)
Investment Type	Equipment or Fixed Capital	• DeLong and Summers (1993) (+,*) • Blomstrom, et al. (1996) (-,_) • Sala-i-Martin (1997a,b) (+,*)
	Non-Equipment	• DeLong and Summers (1991) (+,*)
Labor	Productivity Growth	• Lichtenberg (1992) (+,*)
	Productivity Quality	• Hanushek and Kimko (2000) (+,*)
	Labor Force Part. Rate	• Blomstrom, et al. (1996) (+,*)
Luck	External Debt Dummy	• Easterly, et al. (1993) (-,_)
	External Transfers	• Easterly, et al. (1993) (mixed,_)

	Improvement in Terms of Trade	• Easterly, et al. (1993) (+,*) • Fischer (1993) (+,*) • Barro (1996) (+,*) • Caselli, et al. (1996) (+,*) • Barro (1997) (+,*) • Blattman, et al. (2003) (+,*) (
Money Growth		• Kormendi and Meguire (1985) (+,_)
Neighboring Countries' Education Proxies, Initial Incomes, Investment Ratios and Population Growth Rates		• Ciccone (1996) (*)
Political Instability Proxies		• Barro (1991) (-,*) • Barro and Lee (1994) (-,*) • Sachs and Warner (1995) (-,_) • Alesina, et al. (1996) (-,*) • Caselli, et al. (1996) (-,*) • Easterly and Levine (1997a) (-,*)
Political Rights and Civil Liberties Indices	Civil Liberties	• Kormendi and Meguire (1985) (+,_) • Levine and Renelt (1992) (?,not robust) • Barro and Lee (1994) (-,*)
	Overall	• Sachs and Warner (1995) (+,*)
	Political Rights	• Barro (1991) (?,_) • Barro and Lee (1994) (+,*) • Sala-i-Martin (1997a,b) (+,*)
Political Institutions	Constraints on Executive	• Acemoglu, et al. (2001) (+,*)
	Judicial Independence	• Feld and Voigt (2003) (+,*)
Property Rights	ICRG index	• Knack (1999) (+,*)
	Expropriation Risk	• Acemoglu, et al. (2001) (+,*) • Macarthur and Sachs (2001) (+,*)
Population	Density	• Sachs and Warner (1995) (+,_)
	Growth	• Kormendi and Meguire (1985) (-,*)

		• Levine and Renelt (1992) (-,not robust) • Mankiw, et al. (1992) (-,*) • Barro and Lee (1994) (+,_) • Kelley and Schmidt (1995) (-,*) • Bloom and Sachs (1998) (-,*)
Price Distortions	Consumption Price	• Easterly (1993) (+,_) • Harrison (1996) (-,*)
	Investment Price	• Barro (1991) (-,*) • Easterly (1993) (-,*)
Price Levels	Consumption Price	• Easterly (1993) (+,_)
	Investment Price	• Easterly (1993) (-,*) • Sachs and Warner (1995) (-,*)
Real Exchange Rate	Black Market Premium	• Levine and Renelt (1992) (-,not robust) • Barro and Lee (1994) (-,*) • Barro (1996) (-,*) • Harrison (1996) (-,*) • Easterly and Levine (1997a) (-,*) • Sala-i-Martin (1997a,b) (-,*)
	Distortions	• Dollar (1992) (-,*) • Easterly (1993) (-,_) • Harrison (1996) (-,_) • Sala-i-Martin (1997a,b) (-,*) • Acemoglu, et al. (2002) (-,_)
	Variability	• Dollar (1992) (-,*)
Regional Effects	Absolute Latitude	• Barro (1996) (+,*)
	East Asia Dummy	• Barro and Lee (1994) (+,_) • Barro (1997) (+,_)
	Former Spanish Colonies Dummy	• Barro (1996) (-,*)
	Latin America Dummy	• Barro (1991) (-,*) • Barro and Lee (1994) (-,*) • Barro (1997) (-,_) • Easterly and Levine (1997a) (-,*) • Sala-i-Martin (1997a,b) (-,*)

	Sub-Saharan Africa Dummy	• Barro (1991) (-,*) • Barro and Lee (1994) (-,*) • Barro (1997) (-,_) • Easterly and Levine (1997a) (-,*) • Sala-i-Martin (1997a,b) (-,*)
Religion	Buddhist	• Barro (1996) (+,*)
	Catholic	• Sala-i-Martin (1997a,b) (-,*) • Masters and Sachs (2001) (+,*)
	Confucian	• Barro (1996) (+,*)
	Muslim	• Barro (1996) (+,*) • Sala-i-Martin (1997) (+,*) • Masters and Sachs (2001) (+,_)
	Protestant	• Barro (1996) (+,*) • Sala-i-Martin (1997) (-,*) • Masters and Sachs (2001) (+,*)
	Religious belief	• Barro and McCleary (2003) (+,*)
	Attendance	• Barro and McCleary (2003) (-,*)
Rule of Law Indices		• Barro (1996) (+,*) • Acemoglu, et al. (2001) (+,*) • Easterly and Levine (2001) (-,*) • Dollar and Kraay (2003) (+,_) • Alcala and Ciccone (2004) (+,_/*) • Rodrik et al. (2004) (+,*)
Scale Effects	Total Area	• Barro and Lee (1993) • Sala-i-Martin (1997a,b) (?,_)
	Total Labor force	• Barro and Lee (1993) • Sala-i-Martin (1997a,b) (?,_)
Social Capital and Related	Social "Infrastructure"	• Hall and Jones (1999) (+,*)
	Citizen Satisfaction with Government	• Helliwell and Putnam (2000) (+,*) (within Italy)
	Civic Participation	• Helliwell (1996) (,_) (within Asia)

		• Knack and Keefer (1997) (+,*)
	Groups – as defined by Putnam (1993)	• Keefer and Knack (1997) (-,_),
	Groups - as defined by Olson (1982)	• Keefer and Knack (1997) (+,_),
	Institutional Performance	• Helliwell and Putnam (2000) (+,*) (Italy)
	Civic Community (index of Participation newspaper readership, political behavior)	• Helliwell and Putnam (2000) (+,*) (Italy)
	Trust	• Granato, et al. (1996) (+, *) • Helliwell (1996) (,_)(Asia) • Knack and Keefer (1997) (+,*), • La Porta et al (1997) (+, *) • Beugelsdijk and van Schalk (2001) (,_) • Zak and Knack (2001) (+,*)
	Social Development Index	• Temple and Johnson (1998)
	Extent of Mass Communication	• Temple and Johnson (1998)
	Kinship	• Temple and Johnson (1998)
	Mobility	• Temple and Johnson (1998)
	Middle Class	• Temple and Johnson (1998)
	Outlook	• Temple and Johnson (1998)
	Social capital (WVS)	• Rupasingha, Goetz, and Freshwater (2000) (+,*)
	Social capital (WVS)	• Whiteley (2000) (+,*)
	Social Achievement Norm	• Granato, et al. (1996b) (+,*) • Swank (1996) (-,*)
	Capability	• Temple and Johnson (1998) (+,*)
Trade Policy	Import Penetration	• Levine and Renelt (1992) (?,not

Indices		robust)
	Leamer's Intervention Index	• Levine and Renelt (1992) (-,not robust)
	Years-Open 1950-1990	• Sachs and Warner (1996) (+,*) • Sala-i-Martin (1997a,b) (+,*)
	Openness Indices (growth)	• Harrison (1996) (+,*)
	Openness Indices (level)	• Levine and Renelt (1992) (?,not robust) • Sachs and Warner (1995) (+,*) • Harrison (1996) (+,*) • Wacziarg and Welch (2003) (+,*)
	Outward Orientation	• Levine and Renelt (1992) (?,not robust) • Sala-i-Martin (1997a,b) (?,_)
	Tariff	• Barro and Lee (1994) (-,_) • Sala-i-Martin (1997a,b) (?,_)
Trade Statistics	Fraction of Export/Import/Total-Trade in GDP	• Levine and Renelt (1992) (+,not robust) • Easterly and Levine (1997a) (?,_) • Frankel and Romer (1999) (+,*) • Dollar and Kraay (2003) (+,_) • Alcala and Ciccone (2004) (+,*) • Rodrik et al. (2004) (+,_)
	Fraction of Primary Products in Total Exports	• Sachs and Warner (1996) (-,*) • Sala-i-Martin (1997) (-,*)
	Growth in Export-GDP Ratio	• Feder (1982) (+,*) • Kormendi and Meguire (1985) (+,*) • 20+ studies others
	FDI inflows relative to GDP	• Blomstrom, et al. (1996)
	Machinery and Equipment Import	• Romer (1993) (+,*)
Volatility of Shocks	Growth Innovations	• Kormendi and Meguire (1985) (-,*) • Ramey and Ramey (1995) (-,*)

	Monetary Shock	• Kormendi and Meguire (1985) (-,*)
War	Casualties per Capita	• Easterly, et al. (1993) (-,_)
	Dummy	• Barro and Lee (1994) (-,_) • Easterly and Levine (1997a) (?,_) • Sala-i-Martin (1997a,b) (-,*)
	Duration	• Barro and Lee (1994) (+,_)

.Appendix 3: Instruments Variables for Solow Growth Determinants

Variable	Instrument	Study
GDP growth	Rainfall variation	Miguel, Satyanath, and Sergenti (2003)
GDP – initial	Lagged values	Barro and Sala-i-Martin (2004)
GDP – initial (per capital stock)	Newsprint consumption, and number of radios	Romer (1990)
GDP – initial	Log population initial and trade measure	Bosworth and Collins (2003)
Human Capital	Natural Disasters	Toya, Skidmore, and Robertson (2003)
Investment - Equipment	Equipment prices, WCR survey variables, national savings rates	DeLong and Summers, (1991)
Investment - Education	Age demographics (16) and lagged capital	Cook (2002b)
Investment - Education	Age demographic variables	Higgins (1998)
Investment - Education	Average level	Beaudry, Collard, and Green (2002)
Investment	Initial values of investment/GDP, population growth and GDP	Cho (1996)
Investment	Lagged investment, lagged output, lagged inflation, trade/GDP and gov spend/GDP	Bond, Leblebicioglu, and Schiantarelli (2004).
Investment	Initial investment in sub-period, average savings rate in sub-period,	Beaudry, Collard, and Green (2002)
Population Growth	Initial values of investment/GDP,	Cho (1996)

	population growth and GDP	
Population Growth	Average population growth over sub-period	Beaudry, Collard, and Green (2002)
Neoclassical convergence RHS variables	Civilian fatalities as %of population (and squared), Number of months of occupation by German forces, and number of months of land battles in country	Cook (2002a)

Appendix 4: Instruments Variables for non-Solow Growth Determinants

Variable		Instrument	Study
Capital market imperfections		Degree of insider trading	Bekaert, Harvey, and Lundblad (2001)
Capital Controls		Lagged Values	McKenzie (2001)
Capital Controls		Lagged Values	Grilli and Milesi-Ferretii (1995)
Corruption		Ethnolinguistic Fractionalization	Mauro (1995)
Coups		All variables (some lagged)	Londregan and Poole (1990)
Defense variables		Initial levels of investment, openness, military expenditure and GDP per capita	Guaresma and Reitschuler (2003).
Democracy		Various	Tavares and Wacziarg (2001)
Demography - Urban concentration		Lagged values	Henderson (2000)
Economic freedom		Lagged values	Lundström (2002)
Education	Male and Female level	Religion and civil liberty measures	Dollar and Gatti. (1999)
	Changes in attainment and female/male ratio of change	Change in total fertility rate, educational spend/GDP, initial fertility level	Klasen (2002).
	Change and level	Kyriacou schooling data	Krueger and Lindahl (2001).
Enterprise Size		Legal origin, resource endowments, religious composition, ethnic diversity, and others	Beck, Demirguc-Kunt, and Levine (2003)

Finance	Development	Legal origins and initial income	Demetriades and Law (2004).
	Competition	Legal origin	Claessens and Laeven (2003).
	Various indicators	Initial values of same	King and Levine (1993).
	Depth	"Legal origin" and lagged versions of all explanatory variables	Levine, Loayza, and Beck (2000)
	Depth	Consumption, GDP, and others	Levine and Zervos (1998).
	Depth	Lagged versions of all explanatory variables	Loayza and Ranciere (2002)
	Various "factors"	Wide variety of initial values of regressors and initial inflation	Rousseau and Sylla (2001).
	Depth	Initial values of inflation and financial depth	Rousseau and Wachtel (2002)
Gini Coefficient		Number of municipal townships in 1962, share of labor force in manufacturing in 1990, percentage of revenue from intergovernmental transfers in 1962	Alesina and La Ferrera (2002)
Government Change		Lagged government change and variable reflected composition change in the executive without a government change	Alesina, Ozler, Roubini, and Swagel (1996)
Government Expenditure and Taxation		Various	Agell, Ohlsson, and Thoursie (2003)

Health	Change in Malaria Infection rate	Six variable for % land coverage of type of forest and desert	Gallup, Mellinger, and Sachs (2000).
	Expenditure	Physicians, visits, dialysis, insurance coverage, alcohol, over 65, beds	Rivera and Currais (1999)
Inflation		Lagged explanatory variables	Li and Zou (2002).
Inflation		Initial values of inflation and financial depth	Rousseau and Wachtel (2002)
Infrastructure		Lagged values	Esfahani and Ramirez (2003).
Institutions	Various	Settler mortality rate	Acemoglu, Johnson, and Robinson (2002)
	Various	Historically determined component of current institutional quality	Acemoglu, Johnson, and Robinson (2002).
	Various	Geographically determined component of trade as fraction of GDP AND linguistic origins	Alcalá and Ciccone (2004).
	Various and trust	Lagged values	Keefer and Knack (1997).
	Quality	Mortality rates and initial income)	Demetriades and Law (2004).
Manufacturing Exports		Lagged values	Calderón, Chong, and Zanforlin (2001).
Religiosity		Presence of state religion, regulation of religion, indicator of religious pluralism, and others	Barro and McCleary (2003)

Social Infrastructure		State antiquity	Bockstette, Chanda, and Putterman (2003)
Social Infrastructure		Distance from equator, fraction speaking primary European language, fraction speaking English, Frankel and Romer's log predicted trade share	Hall and Jones (1999).
Stock markets		Lagged stock market activity	Harris (1997)
Technology Gap (first difference)		Lagged (second difference)	Hultberg, Nadiri, and Sickles (2003)
Trade	As Share of GDP	Geographically determined component of trade as fraction of GDP AND linguistic origins	Alcalá and Ciccone (2001).
	Policy indices	Lagged values and others unreported by author	Edwards (1998).
	Policy indices	Lagged values	Amable (2000)
		Geographically determined component of trade as fraction of GDP	Frankel and Romer (1996,1999)
Various - Log initial GDP, broad money to GDP, gov expenditure to GDP		Lagged values	Rousseau (2002)

References

Abramowitz, M., (1986), "Catching Up, Forging Ahead and Falling Behind," *Journal of Economic History*, 46, 385-406.

Acemoglu, D., S. Johnson, and J. Robinson, (2001), "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 91, 5, 1369-1401.

Acemoglu, D., S. Johnson, and J. Robinson, (2002), "Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution," *Quarterly Journal of Economics*, 117, 4, 1231-1294.

Acemoglu, D., S. Johnson, and J. Robinson, (2004), "Institutions as the Fundamental Cause of Long-run Growth," *National Bureau of Economic Research Working Paper no.10481.*

Acemoglu, D., S. Johnson, J. Robinson, and Y. Thaicharoen, (2003), "Institutional Causes, Macroeconomic Symptoms: Volatility, Crises and Growth," *Journal of Monetary Economics*, 50, 1, 49-123.

Acemoglu, D. and F. Zilibotti, (2001), "Productivity Differences," *Quarterly Journal of Economics*, 116, 563-606.

Agell, J., H. Ohlsson, and P. Thoursie, (2003), "Growth Effects of Government Expenditure and Taxation in Rich Countries: A Comment," mimeo, Stockholm University.

Agénor, P.-R., (2004), "Macroeconomic Adjustment and the Poor: Analytical Issues and Cross-Country Evidence," *Journal of Economic Surveys*, 18, 3, 351-408.

Ahn, S. and P. Schmidt, (1995), "Efficient Estimation of Models for Dynamic Panel Data," *Journal of Econometrics*, 68, 5-27.

Aizenman, J. and R. Glick, (2003), "Military Expenditure, Threats and Growth," *National Bureau of Economic Research Working Paper no. 9618*.

Akerlof, G., (1997), "Social Distance and Economic Decisions," *Econometrica*, 65, 5, 1005-1027.

Alcala, F. and A. Ciccone, (2004), "Trade and Productivity," *Quarterly Journal of Economics*, 119, 2, 613-646.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg, (2003), "Fractionalization," *Journal of Economic Growth*, 8, 2, 155-194.

Alesina, A. and E. La Ferrera, (2002), "Who Trusts Others?," *Journal of Public Economics*, 85, 207-235.

Alesina, A., S. Ozler, N. Roubini, and P. Swagel, (1996), "Political Instability and Economic Growth," *Journal of Economic Growth*, 1, 2, 189-211.

Alesina, A. and D. Rodrik, (1994), "Distributive Politics and Economic Growth," *Quarterly Journal of Economics*, 109, 2, 465-490.

Amable, B., (2000), "International Specialisation and Economic Growth," mimeo, University of Lille II.

Anderson, G., (2003), "Making Inferences About the Polarization, Welfare, and Poverty of Nations: A Study of 101 Countries 1970-1995," mimeo, University of Toronto and forthcoming, *Journal of Applied Econometrics*.

Anderson, G., (2004), "Toward an Empirical Analysis of Polarization," *Journal of Econometrics*, 122, 1, 1-26.

Anderson, G. and Y. Ge, (2004), "A New Approach to Convergence: City Types and "Complete" Convergence of Post-Reform Chinese Urban Income Distributions," mimeo, University of Toronto.

Andrade, E., M. Laurini, R. Madalozzo, and P. Valls Pereira, (2004), "Convergence Clubs among Brazilian Municipalities," *Economics Letters*, 83, 179–84.

Andres, J. and A. Lamo, (1995), "Dynamics of the Income Distribution Across OECD Countries," *London School of Economics, Centre for Economic Performance Discussion Paper no. 252.*

Anselin, L., (2001), "Spatial Econometrics," in *A Companion to Theoretical Econometrics*, B. Baltagi, ed., Oxford: Blackwell.

Arellano, M., (2003), *Panel Data Econometrics*, Oxford: Oxford University Press.

Arellano, M. and S. Bond, (1991), "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58, 2, 277-97.

Arellano, M. and O. Bover, (1995), "Another Look at the Instrumental-Variable Estimation of Error-Components Models," *Journal of Econometrics*, 68, 29-51.

Azariadis, C. and A. Drazen, (1990), "Threshold Externalities in Economic Development," *Quarterly Journal of Economics*, 105, 2, 501-526.

Azariadis, C. and J. Stachurski, (2003), "A Forward Projection of the Cross-Country Income Distribution," *Institute of Economic Research, Kyoto University, Discussion Paper No. 570.*

Baltagi, B., S. Song, and W. Koh, (2003), "Testing Panel Data Regression Models with Spatial Error Correlation," *Journal of Econometrics*, 117, 1, 123-50.

Bandyopadhyay, S., (2002), "Polarisation, Stratification, and Convergence Clubs: Some Dynamics and Explanations of Unequal Economic Growth Across Indian States," mimeo, London School of Economics.

Banerjee, A. and E. Duflo, (2003), "Inequality and Growth: What Can the Data Say?," *Journal of Economic Growth*, 8, 3, 267-300.

Barro, R., (1991), "Economic Growth in a Cross Section of Countries," *Quarterly Journal of Economics*, 106, 2, 407-43.

Barro, R., (1996), "Democracy and Growth," *Journal of Economic Growth*, 1, 1, 1-27.

Barro, R., (1997), *Determinants of Economic Growth*, Cambridge: MIT Press.

Barro, R. and J.-W. Lee, (1994), "Sources of Economic Growth (with commentary)," *Carnegie-Rochester Conference Series on Public Policy*, 40, 1-57.

Barro, R., N. G. Mankiw, and X. Sala-i-Martin, (1995), "Capital Mobility in Neoclassical Models of Growth," *American Economic Review*, 85, 1, 103-115.

Barro, R. and R. McCleary, (2003), "Religion and Economic Growth Across Countries," *American Sociological Review*, 68, 5, 760-781.

Barro, R. and X. Sala-i-Martin, (1991), "Convergence Across States and Regions," *Brookings Papers on Economic Activity*, 1, 107-158.

Barro, R. and X. Sala-i-Martin, (1992), "Convergence," *Journal of Political Economy*, 100, 223-51.

Barro, R. and X. Sala-i-Martin, (1995), *Economic Growth*, New York: McGraw-Hill.

Barro, R. and X. Sala-i-Martin, (1997), "Technological Diffusion, Convergence, and Growth," *Journal of Economic Growth*, 2, 1-26.

Barro, R. and X. Sala-i-Martin, (2004), *Economic Growth*, Second edition, Cambridge: MIT Press.

Baumol, W., (1986), "Productivity Growth, Convergence, and Welfare: What the Long-run Data Show," *American Economic Review*, 76, 5, 1072-85.

Beaudry, P., F. Collard, and D. Green, (2002), "Decomposing the Twin-Peaks in the World Distribution of Output-per-Worker," *National Bureau of Economic Research Working Paper no. 9240.*

Beck, T., A. Demirguc-Kunt, and R. Levine, (2003), "Small and Medium Enterprises, Growth, and Poverty: Cross Country Evidence," *World Bank Research Policy Paper 3178*.

Beck, T. and R. Levine, (2004), "Stock Market, Banks, and Growth: Panel Evidence," *Journal of Banking and Finance*, 28, 3, 423-442.

Bekaert, G., C. Harvey, and C. Lundblad, (2001), "Does Financial Liberalization Spur Growth?," *National Bureau of Economic Research Working Paper no. 8245*.

Ben-David, D., (1993), "Equalizing Exchange: Trade Liberalization and Income Convergence," *Quarterly Journal of Economics*, 108, 653-679.

Ben-David, D., (1996), "Trade and Convergence Among Countries," *Journal of International Economics*, 40, 3/4, 279-298.

Bernard, A. and S. Durlauf, (1995), "Convergence in International Output," *Journal of Applied Econometrics*, 10, 2, 97-108.

Bernard, A. and S. Durlauf, (1996), "Interpreting Tests of the Convergence Hypothesis," *Journal of Econometrics*, 71, 1-2, 161-73.

Bernard, A. and C. Jones, (1996), "Comparing Apples to Oranges: Productivity Convergence and Measurement across Industries and Countries," *American Economic Review*, 86, 5, 1216-1238.

Berthelemy, J. and A. Varoudakis, (1996), "Economic Growth, Convergence Clubs, and the Role of Financial Development," *Oxford Economic Papers*, 48, 300-328.

Besley, T. and R. Burgess, (2000), "Land Reform, Poverty Reduction, and Growth: Evidence from India," *Quarterly Journal of Economics*, 115, 2, 389-430.

Besley, T. and R. Burgess, (2002), "The Political Economy of Government Responsiveness: Theory and Evidence from India," *Quarterly Journal of Economics*, 117, 4, 1415-1451.

Besley, T. and R. Burgess, (2004), "Can Labor Regulation Hinder Economic Performance? Evidence from India," *Quarterly Journal of Economics*, 119, 1, 91-134.

Beugelsdijk, S. and T. van Schalk, (2001), "Social Capital and Regional Economic Growth," mimeo, Tilburg University.

Bhargava, A., D. Jamison, L. Lau, and C. Murray, (2001), "Modeling the Effects of Health on Economic Growth," *Journal of Health Economics*, 20, 3, 423-440.

Bianchi, M., (1997), "Testing for Convergence: Evidence from Nonparametric Multimodality Tests," *Journal of Applied Econometrics*," 12, 4, 393-409.

Bils, M. and P. Klenow, (2000), "Does Schooling Cause Growth?," *American Economic Review*, 90, 5, 1160-1183.

Binder, M. and S. Brock, (2004), "A Re-Examination of Determinants of Economic Growth Using Simulatenous Equation Dynamic Panel Data Models," mimeo, Johannes Goethe University, Frankfurt.

Binder, M. and M. H. Pesaran, (1999), "Stochastic Growth Models and their Econometric Implications," *Journal of Economic Growth*, 4, 139-183.

Binder, M. and M. H. Pesaran, (2001), "Life Cycle Consumption Under Social Interactions," *Journal of Economic Dynamics and Control*, 25, 1-2, 35-83.

Blattman, C., J. Hwang, and J. Williamson, (2003), "The Terms of Trade and Economic Growth in the Periphery 1870-1983," *National Bureau of Economic Research Working Paper no. 9940*.

Bliss, C., (1999), "Galton's Fallacy and Economic Convergence," *Oxford Economic Papers*, 51, 4-14.

Bliss, C., (2000), "Galton's Fallacy and Economic Convergence: A Reply to Cannon and Duck," *Oxford Economic Papers*, 52, 420-422.

Blomstrom, M., R. Lipsey, and M. Zejan, (1996), "Is Fixed Investment the Key to Growth?," *Quarterly Journal of Economics*, February, 111, 1, 269-276.

Blonigen, B. and M. Wang, (2004), "Inappropriate Pooling of Wealthy and Poor Countries in Empirical FDI Studies," *National Bureau of Economic Research Working Paper no. 10378*.

Bloom, D., D. Canning, and J. Sevilla, (2002), "Health, Worker Productivity, and Economic Growth," mimeo, Harvard University.

Bloom, D., D. Canning, and J. Sevilla, (2003), "Geography and Poverty Traps," *Journal of Economic Growth*, 8, 355-378.

Bloom, D. and P. Malaney, (1998), "Macroeconomic Consequences of the Russian Mortality Crisis," *World Development*, 26, 11, 2073-2085.

Bloom, D. and J. Sachs, (1998), "Geography, Demography, and Economic Growth in Africa," mimeo, Harvard Institute for International Development.

Bloom, D. and J. Williamson, (1998), "Demographic Transitions and Economic Miracles in Emerging Asia," *World Bank Economic Review*, 12, 3, 419-55.

Blundell, R. and S. Bond, (1998), "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, 87, 1, 115-43.

Bockstette, V., A. Chanda, and L. Putterman, (2003), "States and Markets: The Advantages of an Early Start," *Journal of Economic Growth*, 7, 347-369.

Bond, S., (2002), "Dynamic Panel Data Models: A Guide to Micro Data Methods and Practicem," *Portugese Economic Journal*, 1, 141-162.

Bond, S., A. Hoeffler, and J. Temple, (2001), "GMM Estimation of Empirical Growth Models," *Centre for Economic Policy Research Discussion Paper no. 3048.*

Bond, S., A. Leblebicioglu, and F. Schiantarelli, (2004), "Capital Accumulation and Growth: A New Look at the Empirical Evidence," *Nuffield College, Oxford, Working Paper no. 2004-W8.*

Bosworth, B. and S. Collins, (2003), "The Empirics of Growth: An Update (with discussion)," *Brookings Papers on Economic Activity*, 2, 113-206.

Breiman, L., J. Friedman, R. Olshen and C. Stone, (1984), *Classification and Regression Trees*, Redwood City: Wadsworth Publishing.

Breusch, T. and A. Pagan, (1980), "The Lagrange Multiplier Test and Its Application to Model Specifications in Econometrics," *Review of Economic Studies*, 47, 239-253.

Brock, W. and S. Durlauf, (2001a), "Growth Empirics and Reality," *World Bank Economic Review*, 15, 2, 229-272.

Brock, W. and S. Durlauf, (2001b), "Interactions-Based Models," in *Handbook of Econometrics 5*, J. Heckman and E. Leamer, eds., Amsterdam: North-Holland.

Brock, W., S. Durlauf, and K. West, (2003), "Policy Evaluation in Uncertain Economic Environments (with discussion)," *Brookings Papers on Economic Activity*, 1, 235-322.

Bruno, M. and W. Easterly, (1998), "Inflation Crises and Long-Run Growth," *Journal of Monetary Economics*, 41, 1, 3-26.

Bulli, S., (2001), "Distribution Dynamics and Cross-Country Convergence: New Evidence," *Scottish Journal of Political Economy*, 48, 226-243.

Bun, M. and J. Kiviet, (2001), "The Accuracy of Inference in Small Samples of Dynamic Panel Data Models," *Tinbergen Institute Discussion Paper no. 2001-006/4.*

Calderón, C., A. Chong and L. Zanforlin, (2001), "On the Non-Linearities Between Exports of Manufactures and Economic Growth," *Journal of Applied Economics*, 4, 279-311.

Campos, N. and J. Nugent, (2002), "Who is Afraid of Political Instability?," *Journal of Development Economics*, 67, 157-172.

Cannon, E. and N. Duck, (2000), "Galton's Fallacy and Economic Convergence," *Oxford Economic Papers*, 53, 415-419.

Canova, F., (2004), "Testing for Convergence Clubs in Income Per Capita: A Predictive Density Approach," *International Economic Review*, 45, 1, 49-77.

Canova, F., and A. Marcet, (1995), "The Poor Stay Poor: Non-Convergence Across Countries and Regions," *Centre for Economic Policy Research Discussion Paper 1265.*

Carlino, G. and L. Mills, (1993), "Are U.S. Regional Incomes Converging?: A Time Series Analysis," *Journal of Monetary Economics*, 32, 2, 335-346.

Caselli, F. and W. J. Coleman, (2003), "The World Technology Frontier," mimeo, Harvard University.

Caselli, F., G. Esquivel, and F. Lefort, (1996), "Reopening the Convergence Debate: A New Look at Cross Country Growth Empirics," *Journal of Economic Growth*, 1, 3, 363-89.

Cashin, P., (1995), "Economic Growth and Convergence Across the Seven Colonies of Australasia: 1861-1991," *Economic Record*, 71, 132-144.

Cashin, P. and R. Sahay, (1996), "Regional Economic Growth and Convergence in India," *Finance and Development*, 33, 49-52.

Chesher, A., (1984), "Testing for Neglected Heterogeneity," *Econometrica*, 52, 4, 865-72.

Cho, D., (1996), "An Alternative Interpretation of Conditional Convergence Results," *Journal of Money, Credit and Banking*, 28, 1, 669-681.

Ciccone, A., (1996), "Externalities and Interdependent Growth: Theory and Evidence," mimeo, UC-Berkeley.

Claessens, S. and L. Laeven, (2003), "Competition in the Financial Sector and Growth: A Cross Country Perspective," mimeo, University of Amsterdam.

Cohen, D., (1996), "Tests of the Convergence Hypothesis: Some Further Results," *Journal of Economic Growth*, 1, 3, 351-361.

Collier, P. and J. Gunning, (1999a), "Why Has Africa Grown Slowly?," *Journal of Economic Perspectives*, 13, 3, 3-22.

Collier, P. and J. Gunning, (1999b), "Explaining African Economic Performance," *Journal of Economic Literature*, 37, 1, 64-111.

Conley, T., (1999), "GMM Estimation with Cross-Section Dependence," *Journal of Econometrics*, 92, 1-45.

Conley, T. and E. Ligon, (2002), "Economic Distance and Long-run Growth," *Journal of Economic Growth*, 7, 2, 157-187.

Conley, T. and G. Topa, (2002), "Socio-Economic Distance and Spatial Patterns in Unemployment," *Journal of Applied Econometrics*, 17, 4, 303-327.

Cook, D., (2002a), "World War II and Convergence," *Review of Economics and Statistics*, 84, 1, 131-138.

Cook, D., (2002b), "Education and Growth: Instrumental Variables Estimates," mimeo, Hong Kong University of Science and Technology.

Corrado, L., R. Martin, and M. Weeks, (2004), "Identifying and Intepreting Regional Convergence Clusters Across Europe," *Economic Journal*, forthcoming.

Coulombe, S. and F. Lee, (1995), "Convergence Across Canadian Provinces, 1961 to 1991," *Canadian Journal of Economics*, 28, 886-898.

Dagenais, M., and D. Dagenais, (1997), "Higher Moment Estimators for Linear Regression Models with Errors in the Variables," *Journal of Econometrics*, 76, 1-2, 193-221.

Davidson, R. and J. MacKinnon, (1993), *Estimation and Inference in Econometrics*, Oxford: Oxford University Press.

De Gregorio, J. and P. Guidotti, (1995), "Financial Development and Economic Growth," *World Development*, 23, 3, 433-448.

Deidda, L. and B. Fattouh, (2002), "Non-Linearity Between Finance and Growth," *Economics Letters*, 74, 339-345.

DeLong, J. B., (1988), "Productivity Growth, Convergence, and Welfare: Comment," *American Economic Review*, 78, 5, 1138-1154.

DeLong, J. B. and L. Summers, (1991), "Equipment Investment and Economic Growth," *Quarterly Journal of Economics*, 106, 2, 445-502.

DeLong, J. B. and L. Summers, (1993), "How Strongly do Developing Economies Benefit from Equipment Investment?," *Journal of Monetary Economics*, 32, 3, 395-415.

Demetriades, P. and S. Law, (2004), "Finance, Institutions and Economic Growth," *University of Leicester Working Paper 04/5*.

Desdoigts, A., (1999), "Patterns of Economic Development and the Formation of Clubs," *Journal of Economic Growth*, 4, 3, 305-330.

Diebold, F. and A. Inoue, (2001), "Long Memory and Regime Switching," *Journal of Econometrics*, 105, 1, 131-159.

Dollar, D., (1992), "Outward-Oriented Developing Economies Really Do Grow More Rapidly: Evidence from 95 LDCs, 1976-85," *Economic Development and Cultural Change*, 40, 523-544.

Dollar, D. and R. Gatti, (1999), "Gender Inequality, Income, and Growth: Are Good Times Good for Women?," mimeo, World Bank.

Dollar, D. and A. Kraay, (2002), "Growth *is* Good for the Poor," *Journal of Economic Growth*, 7, 3, 195-225.

Dollar, D. and A. Kraay, (2003), "Institutions, Trade and Growth: Revisiting the Evidence," *Journal of Monetary Economics*, 50, 1, 133-162.

Doppelhofer, G., R. Miller, and X. Sala-i-Martin, (2004), "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," *American Economic Review*, 94, 4, 813-835.

Dowrick, S., (2004), "De-Linearising the Neo-Classical Convergence Model," in *Economic Growth and Macrodynamics: Recent Developments in Economic Theory*, S. Turnovsky, S. Dowrick, and R. Pitchford, eds., New York: Cambridge University Press.

Dowrick, S. and J. Quiggin, (1997), "True Measures of GDP and Convergence," *American Economic Review*, 87, 1, 41-64.

Dowrick, S. and M. Rogers, (2002), "Classical and Technological Convergence: Beyond the Solow-Swan Growth Model," *Oxford Economic Papers*, 54, 369-385.

Draper, D., (1995), "Assessment and Propagation of Model Uncertainty," *Journal of the Royal Statistical Society, series B*, 57, 45-70.

Draper, D., J. Hodges, C. Mallows, and D. Pregibon, (1993), "Exchangeability and Data Analysis (with discussion)," *Journal of the Royal Statistical Society, series A*, 156, 9-37.

Driscoll, J. and A. Kraay, (1998), "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data," *Review of Economics and Statistics*, 80, 4, 549-560.

Duffy, J. and C. Papageorgiou, (2000), "A Cross-Country Empirical Investigation of the Aggregate Production Function Specification," *Journal of Economic Growth*, 5, 87-120.

Durlauf, S., (1996), "Controversy on the Convergence and Divergence of Growth Rates," *Economic Journal*, 106, 1016-1018.

Durlauf, S., (2001), "Manifesto for a Growth Econometrics," *Journal of Econometrics*, 100, 1, 65-69.

Durlauf, S., (2002), "On the Empirics of Social Capital," *Economic Journal*, 112, 459-479.

Durlauf, S. and P. Johnson, (1995), "Multiple Regimes and Cross Country Growth Behaviour," *Journal of Applied Econometrics*, 10, 4, 365-84.

Durlauf, S., A. Kourtellos, and A. Minkin, (2001), "The Local Solow Growth Model," *European Economic Review*, 45, 4-6, 928-40.

Durlauf, S. and D. Quah, (1999), "The New Empirics of Economic Growth," in *Handbook of Macroeconomics*, J. Taylor and M. Woodford, eds., Amsterdam: North Holland.

Easterly, W., (1993), "How Much Do Distortions Affect Growth?," *Journal of Monetary Economics*, 32, 2, 187-212.

Easterly, W., (1994), "Economic Stagnation, Fixed Factors, and Policy Thresholds," *Journal of Monetary Economics*, 33, 3, 525-57.

Easterly, W., (1996), "When is Stabilization Expansionary?," *Economic Policy*, 22, 67-98.

Easterly, W., (2001), "The Lost Decades: Developing Countries' Stagnation in Spite of Policy Reform 1980-1998," *Journal of Economic Growth*, 6, 2, 135-57.

Easterly, W., M. Kremer, L. Pritchett, and L. Summers, (1993), "Good Policy or Good Luck? Country Growth Performance and Temporary Shocks," *Journal of Monetary Economics*, 32, 459-483.

Easterly, W. and R. Levine, (1997a), "Africa's Growth Tragedy: Policies and Ethnic Divisions," *Quarterly Journal of Economics*, 112, 4, 1203-50.

Easterly, W. and R. Levine, (1997b), "Troubles with the Neighbours: Africa's Problem, Africa's Opportunity," *Journal of African Economies*, 7, 1, 120-42.

Easterly, W. and R. Levine, (2001), "It's Not Factor Accumulation: Stylized Facts and Growth Models," *World Bank Economic Review*, 15, 177-219.

Eaton, J. and S. Kortum, (1999), "International Technology Diffusion: Theory and Measurement," *International Economic Review*, 40, 3, 537-570.

Eaton, J. and S. Kortum, (2001), "Trade in Capital Goods," *European Economic Review*, 45 7, 1195-1235.

Edwards, S., (1998), "Openness, Productivity and Growth: What Do We Really Know?," *Economic Journal*, 108, 383-98.

Edwards, S. and I. Magendzo, (2003), "Strict Dollarization and Economic Performance: An Empirical Investigation," *International Journal of Finance and Economics*, 8, 4, 351-363.

Eichengreen, B. and D. Leblang, (2003), "Capital Account Liberalization and Growth: Was Mr. Mahathir Right?," *International Journal of Finance and Economics*, 8, 3, 205-224.

Eicker, F., (1967), "Limit Theorems for Regressions with Unequal and Dependent Errors," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley: University of California.

Esfahani, H. and M. Ramirez, (2003), "Institutions, Infrastructure, and Economic Growth," *Journal of Development Economics*, 70, 2, 443-477.

Evans, P., (1996), "Using Cross-Country Variances to Evaluate Growth Theories," *Journal of Economic Dynamics and Control*, 20, 1027-1049.

Evans, P., (1997), "How Fast Do Economies Converge?," *Review of Economics and Statistics*, 79, 2, 219-225.

Evans, P., (1998), "Using Panel Data to Evaluate Growth Theories," *International Economic Review*, 39, 2, 295-306.

Feder, G., (1982), "On Exports and Economic Growth," *Journal of Development Economics*, 12, 1, 59-74.

Feld, L. and S. Voigt, (2003), "Economic Growth and Judicial Independence: Cross Country Evidence Using a New Set of Indicators," *European Journal of Political Economy*, 19, 3, 497-527.

Fernandez, C., E. Ley and M. Steel, (2001a), "Model Uncertainty in Cross-Country Growth Regressions," *Journal of Applied Econometrics*, 16, 5, 563-76.

Fernandez, C., E. Ley and M. Steel, (2001b) "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100, 2, 381-427.

Feyrer, J., (2003), "Convergence By Parts," mimeo, Dartmouth College.

Fiaschi, D. and A. Lavezzi, (2004), "Distribution Dynamics and Nonlinear Growth," *Journal of Economic Growth*, 8, 379-401.

Fischer, S., (1993), "The Role of Macroeconomic Factors in Growth," *Journal of Monetary Economics*, 32, 3, 485-512.

Forbes, K., (2000), "A Reassessment of the Relationship between Inequality and Growth," *American Economic Review*, 90, 4, 869-87.

Frankel, J., (2003), "Discussion," *Brookings Papers on Economic Activity*, 2, 189-199.

Frankel, J. and D. Romer, (1996), "Trade and Growth: An Empirical Investigation," *National Bureau of Economic Research Working Paper no. 5476.*

Frankel, J. and D. Romer, (1999), "Does Trade Cause Growth?," *American Economic Review*, 89, 3, 379-399.

Frankel, J., D. Romer, and T. Cyrus, (1996), "Trade and Growth in East Asian Countries: Cause and Effect?," *National Bureau of Economic Research Working Paper no. 5732.*

Friedman, J., (1987), "Exploratory Projection Pursuit," *Journal of the American Statistical Association*, 82, 249-266.

Friedman, J. and J. Tukey, (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Transactions on Computers*, C23, 881-890.

Friedman, M., (1992), "Do Old Fallacies Ever Die?," *Journal of Economic Literature*, 30, 2129-2132.

Galor, O., (1996), "Convergence? Inferences from Theoretical Models," *Economic Journal*, 106, 1056-1069.

Gallup, J., A. Mellinger, and J. Sachs, (2000), "The Economic Burden of Malaria," *Harvard University, Center for International Development Working Paper No. 1.*

Giavazzi, F. and G. Tabellini, (2004), "Economic and Political Liberalizations," *Centre for Economic Policy Research Discussion Paper no. 4579.*

Goetz, S. and D. Hu, (1996), "Economic Growth and Human Capital Accumulation: Simultaneity and Expanded Convergence Tests," *Economic Letters*, 51, 355-362.

Graham, B. and J. Temple, (2003), "Rich Nations, Poor Nations: How Much Can Multiple Equilibria Explain?," mimeo, University of Bristol.

Granato, J., R. Inglehart, and D. Leblang, (1996), "The Effect of Cultural Values on Economic Development: Theory, Hypotheses, and Some Empirical Tests," *American Journal of Political Science*, 40, 3, 607-31.

Granger, C. W. J., (1980), "Long-Memory Relationships and the Aggregation of Dynamic Models," *Journal of Econometrics*, 14, 227-238.

Greasley, D. and L. Oxley, (1997), "Time-Series Tests of the Convergence Hypothesis: Some Positive Results," *Economics Letters*, 56, 143-147.

Grier, K. and G. Tullock, (1989), "An Empirical Analysis of Cross National Economic Growth, 1951-80," *Journal of Monetary Economics*, 24, 2, 259-76.

Grilli, V. and G. Milesi-Ferretii, (1995), "Economic Effects and Structural Determinants of Capital Controls," *International Monetary Fund Working Paper WP/95/31*.

Guaresma, J. and G. Reitschuler, (2003), "'Guns or Butter?' Revisited: Robustness and Nonlinearity Issues in the Defense-Growth Nexus," mimeo, University of Vienna.

Hahn, J., (1999), "How Informative Is the Initial Condition in the Dynamic Panel Model with Fixed Effects?," *Journal of Econometrics*, 93, 2, 309-26.

Hahn, J., J. Hausman, and G. Kuersteiner, (2001), "Bias Corrected Instrumental Variables Estimation for Dynamic Panel Models with Fixed Effects," mimeo, Massachusetts Institute for Technology.

Hall, A., (1987), "The Information Matrix Test for the Linear Model," *Review of Economic Studies*, 54, 2, 257-63.

Hall, R., and C. Jones, (1999), "Why Do Some Countries Produce So Much More Output Per Worker Than Others?," *Quarterly Journal of Economics*, 114, 1, 83-116.

Hall, S., D. Robertson, and M. Wickens, (1997), "Measuring Economic Convergence," *International Journal of Finance and Economics*, 2, 131-143.

Hamoudi, A. and J. Sachs, (2000), "Economic Consequences of Health Status: A Review of the Evidence," *Harvard University, CID Working Paper No. 30*.

Hansen, B., (2000), "Sample Splitting and Threshold Estimation," *Econometrica*, 68, 3, 575-603.

Hanushek, E. and D. Kimko, (2000), "Schooling, Labor-Force Quality, and the Growth of Nations," *American Economic Review*, 90, 5, 1184-1208.

Harberger, A., (1987), "Comment," in *Macroeconomics Annual 1987*, S. Fischer, ed., Cambridge: MIT Press.

Harris, D., (1997), "Stock Markets and Development: A Re-Assessment," *European Economic Review*, 41, 1, 139-146.

Harrison, A., (1996), "Openness and Growth: A Time-Series, Cross-Country Analysis for Developing Countries," *Journal of Development Economics*, 48, 2, 419-447.

Hausmann, R., L. Pritchett and D. Rodrik, (2004), "Growth Accelerations," *Centre for Economic Policy Research Discussion Paper no. 4538.*

Helliwell, J., (1996), "Economic Growth and Social Capital in Asia," in *The Asia Pacific Region in the Global Economy: A Canadian Perspective*, R. Harris, ed., Calgary: University of Calgary Press.

Helliwell, J. and R. Putnam, (2000), "Economic Growth and Social Capital in Italy," in *Social Capital: A Multifaceted Perspective*, P. Dasgupta and I. Seragilden, eds., Washington DC: World Bank.

Henderson, D. and R. Russell, (2004), "Human Capital and Convergence: A Production Frontier Approach," mimeo, SUNY Binghamton and forthcoming, *International Economic Review.*

Henderson, J. V., (2000), "The Effects of Urban Concentration on Economic Growth," *National Bureau of Economic Research Working Paper no. 7503*.

Hendry, D., (1995), *Dynamic Econometrics*, New York: Oxford University Press.

Hendry, D. and H.-M. Krolzig, (2004), "We Ran One Regression," mimeo, Oxford University.

Hendry, D. and H.-M. Krolzig, (2005), "The Properties of Automatic Gets Modelling," *Economic Journal*, forthcoming.

Henry, P., (2000), "Do Stock Market Liberalizations Cause Investment Booms?," *Journal of Financial Economics*, 58, 1-2, 301-334.

Henry, P., (2003), "Capital Account Liberalization, the Cost of Capital, and Economic Growth," *American Economic Review*, 93, 2, 91-96.

Heston, A., R. Summers and B. Aten, (2002), "Penn World Table Version 6.1," *Center for International Comparisons at the University of Pennsylvania (CICUP).*

Higgins, M., (1998), "Demography, National Savings and International Capital Flows," *International Economic Review*, 39, 343-369.

Hobijn, B. and Franses, P. H., (2000), "Asymptotically Perfect and Relative Convergence of Productivity," *Journal of Applied Econometrics*, 15, 59-81.

Hoeffler, A., (2002), "The Augmented Solow Model and the African Growth Debate," *Oxford Bulletin of Economics and Statistics*, 64, 2, 135-158.

Holtz-Eakin, D., W. Newey, and H. Rosen, (1988), "Estimating Vector Autoregressions with Panel Data," *Econometrica*, 56, 6, 1371-1395.

Hoover, K. and S. Perez, (2004), "Truth and Robustness in Cross-Country Growth Regressions," *Oxford Bulletin of Economics and Statistics*, forthcoming.

Howitt, P., (2000), "Endogenous Growth and Cross-Country Income Differences," *American Economic Review*, 90, 4, 829-846.

Hultberg, P., M. Nadiri, and R. Sickles, (2003), "Cross-country Catch-up in the Manufacturing Sector: Impacts of Heterogeneity on Convergence and Technology adoption," mimeo, University of Wyoming.

Hulten, C., (1996), "Infrastructure Capital and Economic Growth: How Well You Use It May Be More Important Than How Much You Have," *National Bureau of Economic Research Working Paper no. 5847*.

Islam, N., (1995), "Growth Empirics: A Panel Data Approach," *Quarterly Journal of Economics*, 110, 4, 1127-70.

Islam, N. (1998), "Growth Empirics: A Panel Data Approach-A Reply," *Quarterly Journal of Economics*, 113, 325-329.

Islam, N., (2003), "What Have We Learned from the Convergence Debate?," *Journal of Economic Surveys*, 17, 309-362.

Johnson, P., (2000), "A Nonparametric Analysis of Income Convergence Across the US States," *Economics Letters*, 69, 219-223.

Johnson, P., (2004), "A Continuous State Space Approach to 'Convergence by Parts'," mimeo, Vassar College, and forthcoming, *Economic Letters*.

Johnson, P. and L. Takeyama, (2001), "Initial Conditions and Economic Growth in the US States," *European Economic Review,* 45, 4-6, 919-27.

Jones, C., (1995), "Time Series Tests of Endogenous Growth Models," *Quarterly Journal of Economics*, 110, 2, 495-525.

Jones, C., (1997), "Convergence Revisited," *Journal of Economic Growth*, 2, 2, 131-53.

Jones, L. and R. Manuelli, (1990), "A Convex Model of Equilibrium Growth: Theory and Policy Implications," *Journal of Political Economy*, 98, 5, 1008-1038.

Judson, R. and A. Owen, (1999), "Estimating Dynamic Panel Data Models: A Guide for Macroeconomists," *Economics Letters*, 65, 9-15.

Kalaitzidakis, P., T. Mamuneas, and T. Stengos, (2000), "A Non-linear Sensitivity Analysis of Cross Country Growth Regressions," *Canadian Journal of Economics*, 33, 3, 604-17.

Kaufmann, D., A. Kraay, and M. Mastruzzi, (2003), "Governance Matters III: Governance Indicators for 1996-2002," mimeo, World Bank.

Kaufmann, D., A. Kraay, and P. Zoido-Lobaton, (1999a), "Aggregating Governance Indicators," *World Bank Policy Research Department Working Paper No. 2195*.

Kaufmann, D., A. Kraay, and P. Zoido-Lobaton, (1999b), "Governance Matters," *World Bank Policy Research Department Working Paper No. 2196.*

Keefer, P. and S. Knack, (1997), "Why Don't Poor Countries Catch Up? A Cross-National Test of an Institutional Explanation," *Economic Inquiry*, 35, 3, 590-602.

Kelley, A. and R. Schmidt, (1995), "Aggregate Population and Economic Growth Correlations: The Role of the Components of Demographic Change," *Demography*, 32, 543-555.

Kelly, M., (1992), "On Endogenous Growth with Productivity Shocks," *Journal of Monetary Economics*, 30, 1, 47-56.

Kelly, T., (1997), "Public Expenditures and Growth," *Journal of Development Studies*, 34, 1, 60-84.

King, R. and R. Levine, (1993), "Finance and Growth: Schumpeter Might be Right," *Quarterly Journal of Economics*, 108, 3, 717-737.

Kiviet, J., (1995), "On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models," *Journal of Econometrics*, 68, 1, 53-78

Kiviet, J., (1999), "Expectations of Expansions for Estimators in a Dynamic Panel Data Model: Some Results for Weakly Exogenous Regressors," in *Analysis of Panels and Limited Dependent Variable Models: In Honour of G. S. Maddala*, C. Hsiao, ed., Cambridge: Cambridge University Press.

Klasen, S., (2002), "Lower Schooling for Girls, Slower Growth for All? Cross-Country Evidence on the Effect of Gender Inequality in Education on Economic Development," *World Bank Economic Review*, 16, 3, 345-373.

Klenow, P. and A. Rodriguez-Clare, (1997a), "The Neoclassical Revival in Growth Economics: Has it Gone Too Far?," in *Macroeconomics Annual 1997*, B. Bernanke and J. Rotemberg, eds., Cambridge: MIT Press.

Klenow, P. and A. Rodriguez-Clare, (1997b), "Economic Growth: A Review Essay," *Journal of Monetary Economics*, 40, 597-617.

Klepper, S., (1988), "Regression Diagnostics for the Classical Errors-in-Variables Model," *Journal of Econometrics*, 37, 225-250.

Klepper, S. and E. Leamer, (1984), "Consistent Sets of Estimates for Regressions with Errors in All Variables," *Econometrica*, 52, 1, 163-83.

Knack, S., (1999), "Social Capital, Growth and Poverty: A Survey of Cross Country Evidence," *World Bank, Social Capital Initiative Working Paper No. 7*.

Knack, S. and P. Keefer, (1995), "Institutions and Economic Performance: Cross Country Tests Using Alternative Institutional Measures," *Economics and Politics*, 7, 3, 207-27.

Knack, S. and Keefer, P., (1997), "Does Social Capital Have an Economic Payoff? A Cross-Country Investigation," *Quarterly Journal of Economics*, 112, 4, 1252-1288.

Knight, M., N. Loayza, and D. Villaneuva, (1993), "Testing the Neoclassical Growth Model," *IMF Staff Papers*, 40, 512-541.

Knowles, S., (2001), "Inequality and Economic Growth: The Empirical Relationship Reconsidered in the Light of Comparable Data," *CREDIT Research Paper 01/03*.

Knowles, S. and P. Owen, (1995), "Health Capital and Cross-Country Variation in Income per Capita in the Mankiw-Romer-Weil Model," *Economics Letters*, 48, 1, 99-106.

Kocherlakota, N. and K.-M. Yi, (1997), "Is There Endogenous Long-Run Growth? Evidence from the United States and the United Kingdom," *Journal of Money, Credit and Banking*, 29, 2, 235-262.

Kormendi, R. and P. Meguire, (1985), "Macroeconomic Determinants of Growth: Cross Country Evidence," *Journal of Monetary Economics*, 16, 2, 141-63.

Kourtellos, A., (2003a), "Modeling Parameter Heterogeneity in Cross-Country Growth Regression Models," mimeo, University of Cyprus.

Kourtellos, A., (2003b), "A Projection Pursuit Approach to Cross-Country Growth Data," mimeo, University of Cyprus.

Krasker, W. and J. Pratt, (1986), "Bounding the Effects of Proxy Variables on Regression Coefficients," *Econometrica*, 54, 3, 641-55

Krasker, W. and J. Pratt, (1987), "Bounding the Effects of Proxy Variables on Instrumental Variables Coefficients," *Journal of Econometrics*, 35, 2/3, 233-52.

Kremer, M., A. Onatski, and J. Stock, (2001), "Searching for Prosperity," *Carnegie-Rochester Conference Series on Public Policy*, 55, 275-303.

Krueger, A. and M. Lindahl, (2001), "Education for Growth: Why and for Whom," *Journal of Economic Literature*, 39, 4, 1101-1136.

Lamo, A., (2000), "On Convergence Empirics: Some Evidence for Spanish Regions," *Investigaciones Economicas*, 24, 681-707.

Landes, D., (1998), *The Wealth and Poverty of Nations: Why Some are So Rich and Some So Poor*, New York: W. W. Norton and Company.

LaPorta, R., F. Lopez-de-Silanes, A. Shleifer, and R. Vishny, (1997), "Trust in Large Organizations," *American Economic Review*, 87, 2, 333-8.

Leamer, E., (1983), "Let's Take the Con Out of Econometrics," *American Economic Review*, 73, 1, 31-43.

Leamer, E. and H. Leonard, (1983), "Reporting the Fragility of Regression Estimates," *Review of Economics and Statistics*, 65, 2, 306-17.

Lee, K., M. Pesaran, and R. Smith, (1997), "Growth and Convergence in Multi Country Empirical Stochastic Solow Model," *Journal of Applied Econometrics*, 12, 4, 357-92.

Lee, K., M. Pesaran, and R. Smith, (1998), "Growth Empirics: A Panel Data Approach: A Comment," *Quarterly Journal of Economics*, 113, 1, 319-23.

Levine, R., N. Loayza, and T. Beck, (2000), "Financial Intermediation and Growth: Causality and Causes," *Journal of Monetary Economics*, 46, 1, 31-77.

Levine, R. and D. Renelt, (1991), "Cross-Country Studies of Growth and Policy: Methodological, Conceptual, and Statistical Problems." *World Bank PRE Working Paper no. 608*.

Levine, R. and D. Renelt, (1992), "A Sensitivity Analysis of Cross-Country Growth Regressions," *American Economic Review*, 82, 4, 942-63.

Levine, R. and S. Zervos, (1993), "What We Have Learned About Policy and Growth from Cross-Country Regressions," *American Economic Review*, 83, 2, 426-430.

Levine, R. and S. Zervos, (1998), "Stock Markets, Banks and Economic Growth," *American Economic Review*, 88, 3, 537-558.

Li, H. and H.-F. Zou, (2002), "Inflation, Growth, and Income Distribution: A Cross Country Study," *Annals of Economics and Finance*, 3, 1, 85-101.

Li, Q. and D. Papell, (1999), "Convergence of International Output: Time Series Evidence for 16 Countries," *International Review of Economics and Finance*, 8, 267-280.

Lichtenberg, F., (1992), "R&D Investment and International Productivity Differences," *National Bureau of Economic Research Working Paper no. 4161*.

Liu, Z. and T. Stengos, (1999), "Non-Linearities in Cross Country Growth Regressions: A Semiparametric Approach," *Journal of Applied Econometrics*, 14, 5, 527-38.

Loayza, N. and R. Ranciere, (2002), "Financial Development, Financial Fragility, and Growth," *CESifo Working Paper Series no. 684.*

Loewy, M. and D. Papell, (1996), "Are U.S. Regional Incomes Converging? Some Further Evidence," *Journal of Monetary Economics*, 38, 3, 587-598.

Loh, W.-Y., (2002), "Regression Trees with Unbiased Variable Selection and Interaction Detection," *Statistica Sinica*, 12, 361-386.

Londregan, J. and K. Poole, (1990), **"**Poverty, the Coup Trap, and the Seizure of Executive Power," *World Politics*, 4, 2, 151-183.

Long, J. and L. Ervin, (2000), "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model," *American Statistician*, 54, 3, 217-224.

Lucas, R., (1988), "On the Mechanics of Economic Development," *Journal of Monetary Economics*, 22, 1, 3-42.

Lundström, S., (2002), "On Institutions, Economic Growth, and the Environment," mimeo, Goteborg University.

Maasoumi, E., J. Racine, and T. Stengos, (2003), "Growth and Convergence: A Profile of Distribution Dynamics and Mobility," mimeo, Southern Methodist University.

MacKinnon, J. and H. White, (1985), "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305-325.

Maddala, G. and S. Wu, (2000), "Cross-Country Growth Regressions: Problems of Heterogeneity, Stability, and Interpretation," *Applied Economics*, 32, 635-642.

Maddison, A., (1982), *Phases of Capitalist Development*: New York: Oxford University Press.

Maddison, A., (1989), *The World Economy in the 20th Century*, OECD: Paris.

Mamuneas, T., A. Savvides, and T. Stengos, (2004), "Economic Development and Return to Human Capital: A Smooth Coefficient Semiparametric Approach," mimeo, University of Guelph and forthcoming, *Journal of Applied Econometrics*.

Mankiw, N. G., (1995), "The Growth of Nations," *Brookings Papers on Economic Activity*, 1, 275-310.

Mankiw, N. G., D. Romer, and D. Weil, (1992), "A Contribution to the Empirics of Economic Growth," *Quarterly Journal of Economics*, 107, 2, 407-37.

Manski, C., (1993), "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 60, 3, 531-42.

Masanjala, W. and C. Papageorgiou, (2004), "Rough and Lonely Road to Prosperity: A Reexamination of the Sources of Growth in Africa Using Bayesian Model Averaging," mimeo, Louisiana State University.

Masters, W. and M. McMillan, (2001), "Climate and Scale in Economic Growth," *Journal of Economic Growth*, 6, 3, 167-186.

Masters, W. and J. Sachs, (2001), "Climate and Development," mimeo, Purdue University.

Mauro, P., (1995), "Corruption and Growth," *Quarterly Journal of Economics*, 110, 3, 681-713.

McArthur, J. and J. Sachs, (2001), "Institutions and Geography: Comment on Acemoglu, Johnson and Robinson," *National Bureau of Economic Research Working Paper no. 8114*.

McCarthy, D., H. Wolf and Y. Wu, (2000), "The Growth Costs of Malaria," *National Bureau of Economic Research Working Paper no. 7541*.

McKenzie, D., (2001), "The Impact of Capital Controls on Growth Convergence," *Journal of Economic Development*, 26, 1, 1-24.

Michelacci, C. and P. Zaffaroni, (2000), "Fractional (Beta) Convergence," *Journal of Monetary Economics*, 45, 129-153.

Miguel, E., S. Satyanath, and E. Segenti, (2003), "Economic Shocks and Civil Conflict: An Instrumental Variables Approach," mimeo, UC Berkeley.

Minier, J., (1998), "Democracy and Growth: Alternative Approaches," *Journal of Economic Growth*, 3, 3, 241-266.

Mokyr, J., (1992), *Lever of Riches: Technological Creativity and Economic Progress*, Princeton: Princeton University Press.

Motley, B., (1998). "Growth and Inflation: A Cross-Country Study." *FRBSF Economic Review*, 1.

Murphy, K., A. Shleifer, and R. Vishny, (1991), "The Allocation of Talent: Implications for Growth," *Quarterly Journal of Economics*, 106, 2, 503-530.

Nelson, M. and R. Singh, (1994), "The Deficit-Growth Connection: Some Recent Evidence from Developing Countries," *Economic Development and Cultural Change*, 42, 167-191.

Nerlove, M., (1999), "Properties of Alternative Estimators of Dynamic Panel Models: An Empirical Analysis of Cross-Country Data for the Study of Economic Growth," in *Analysis of Panels and Limited Dependent Variable Models: In Honour of G. S. Maddala*, C. Hsiao, ed., Cambridge: Cambridge University Press.

Nerlove, M., (2000), "Growth Rate Convergence, Fact or Artifact? An Essay on Panel Data Econometrics," in *Panel Data Econometrics: Future Directions: Papers in Honour of Professor Pietro Balestra*, E. Ronchetti, ed., Amsterdam: North-Holland.

Nickell, S., (1981), "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 49, 6, 1417-1426.

Odedokun, M., (1996), "Alternative Econometric Approaches for Analysing the Role of the Financial Sector in Economic Growth: Time Series Evidence from LDCs," *Journal of Development Economics* 50, 1, 119-146.

Olson, M., (1982), *The Rise and Decline of Nations*, New Haven: Yale University Press.

Paap, R. and H. van Dijk, (1998), "Distribution and Mobility of Wealth of Nations," *European Economic Review*, 42, 7, 1269-93.

Papageorgiou, C., (2002), "Trade as a Threshold Variable for Multiple Regimes," *Economics Letters*, 71, 1, 85-91.

Papageorgiou, C. and W. Masanjala, (2004), "The Solow Model with CES Technology: Nonlinearities with Parameter Heterogeneity," *Journal of Applied Econometrics*, 19, 2, 171-201.

Perron, P., (1989), "The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis," *Econometrica*, 57, 6, 1361-1401.

Persson, J., (1997), "Convergence Across Swedish Counties, 1911-1993," *European Economic Review*, 41, 1835-1852.

Persson, T. and G. Tabellini, (1994), "Is Inequality Harmful for Growth?," *American Economic Review*, 84, 3, 600-621

Persson, T. and G. Tabellini, (2003), *The Economic Effects of Constitutions*, Cambridge: MIT Press.

Pesaran, M. H., (2004a), "A Pair-Wise Approach to Testing for Output and Growth Convergence," mimeo, University of Cambridge.

Pesaran, M. H., (2004b), "General Diagnostic Tests for Cross-Section Dependence in Panels," mimeo, University of Cambridge.

Pesaran, M. H., Y. Shin, and R. Smith, (1999), "Pooled Mean Group Estimation of Dynamic Heterogeneous Panels," *Journal of the American Statistical Association*, 94, 446, 621-34.

Pesaran, M. H. and R. Smith, (1995), "Estimating Long-run Relationships from Dynamic Heterogeneous Panels," *Journal of Econometrics*, 68, 1, 79-113.

Phillips, P. and D. Sul, (2002), "Dynamic Panel Estimation and Homogeneity Testing Under Cross-Section Dependence," *Cowles Foundation Discussion Paper no. 1362.*

Phillips, P. and D. Sul, (2003), "The Elusive Empirical Shadow of Growth Convergence," *Cowles Foundation Discussion Paper no. 1398.*

Prescott, E., (1998), "Needed: A Theory of Total Factor Productivity," *International Economic Review*, 39, 525-551.

Pritchett, L., (1997), "Divergence, Big Time," *Journal of Economic Perspectives*, 11, 3, 3-17.

Pritchett, L., (2000a), "Understanding Patterns of Economic Growth: Searching for Hills among Plateaus, Mountains, and Plains," *World Bank Economic Review*, 14, 2, 221-50.

Pritchett, L., (2000b), "The Tyranny of Concepts: CUDIE (Cumulated, Depreciated, Investment Effort) Is Not Capital," *Journal of Economic Growth*, 5, 4, 361-84.

Putnam, R., R. Leonardi and R. Nanetti, (1993), *Making Democracy Work*, Princeton: Princeton University Press.

Rupasingha, A., S. Goetz, and D. Freshwater, (2000), "Social Capital and Economic Growth: A County-Level Analysis," *Journal of Agricultural and Applied Economics*, 32, 3, 565-72.

Quah, D., (1993a), "Galton's Fallacy and Tests of the Convergence Hypothesis," *Scandinavian Journal of Economics*, 95, 427-443.

Quah, D., (1993b), "Empirical Cross-Section Dynamics in Economic Growth," *European Economic Review*, 37, 2-3, 426-34.

Quah, D., (1996a), "Twin Peaks: Growth and Convergence in Models of Distribution Dynamics," *Economic Journal*, 106, 437, 1045-55.

Quah, D., (1996b), "Empirics for Economic Growth and Convergence," *European Economic Review*, 40, 6, 1353-75.

Quah, D., (1996c), "Convergence Empirics Across Economies with (Some) Capital Mobility," *Journal of Economic Growth*, 1, 1, 95-124.

Quah, D., (1997), "Empirics for Growth and Distribution: Stratification, Polarization, and Convergence Clubs," *Journal of Economic Growth*, 2, 1, 27-59.

Quah, D., (2001), "Searching for Prosperity: A Comment," *Carnegie-Rochester Conference Series on Public Policy*, 55, 305-19.

Ram, R., (1999), "Financial Development and Economic Growth," *The Journal of Development Studies*, 27, 2, 151-167.

Ramey, G. and V. Ramey, (1995), "Cross-Country Evidence on the Link Between Volatility and Growth," *American Economic Review*, 85, 5, 1138-1151.

Reichlin, L., (1999), "Discussion of 'Convergence as Distribution Dynamics', by Danny Quah," in *Market Integration, Regionalism, and the Global Economy*, R. Baldwin, D. Cohen, A. Sapir, and A. Venables, eds., Cambridge: Cambridge University Press.

Rivera, B. and L. Currais, (1999), "Economic Growth and Health: Direct Impact or Reverse Causation?," *Applied Economics Letters*, 6, 11, 761-64.

Robertson, D and J. Symons, (1992), "Some Strange Properties of Panel Data Estimators," *Journal of Applied Econometrics*, 7, 2, 175-89.

Rodriguez, F. and D. Rodrik, (2001), "Trade Policy and Economic Growth: A User's Guide," in *Macroeonomics Annual 2000*, B. Bernanke and K. Rogoff, eds., Cambridge: MIT Press.

Rodrik, D., (1999), "Where Did All the Growth Go? External Shocks, Social Conflict, and Growth Collapses," *Journal of Economic Growth*, 4, 4, 385-412.

Rodrik, D. (ed.), (2003), *In Search of Prosperity: Analytic Narratives on Economic Growth*, Princeton: Princeton University Press.

Rodrik, D., A. Subramanian, and F. Trebbi, (2004), "Institutions Rule: The Primacy of Institutions Over Geography and Integration in Economic Development," *Journal of Economic Growth*, 9, 2, 131-165.

Romer, D., (2001), *Advanced Macroeconomics*, New York: McGraw-Hill.

Romer, P., (1986), "Increasing Returns and Long-run Growth," *Journal of Political Economy*, 94, 5, 1002-1037.

Romer, P., (1990), "Human Capital and Growth: Theory and Evidence," *Carnegie-Rochester Series on Public Policy*, 32, 251-286.

Romer, P., (1993), "Idea Gaps and Object Gaps in Economic Development," *Journal of Monetary Economics*, 32, 3, 543-573.

Roubini, N. and X. Sala-i-Martin, (1992), "Financial Repression and Economic Growth," *Journal of Development Economics*, 39, 5-30.

Roubini, N. and X. Sala-i-Martin, (1992), "Financial Repression and Economic Performance: Historical Evidence from Five Industrialized Countries," *Journal of Money, Credit and Banking*, 30, 4, 657-678.

Rousseau, P., (2002), "Historical Perspectives on Financial Development and Economic Growth," *Review Federal Reserve Bank of St. Louis*, 84, 4.

Rousseau, P. and R. Sylla, (2001), "Financial Systems, Economic Growth, and Globalization," mimeo, Vanderbilt University.

Rousseau, P. and P. Wachtel, (2002), "Inflation Thresholds and the Finance-Growth Nexus," *Journal of International Money and Finance*, 21, 6, 77-793.

Sachs, J., (2003), "Institutions Don't Rule: Direct Effects of Geography on Per Capita Income," *National Bureau of Economic Research Working Paper no. 9490*.

Sachs, J. and A. Warner, (1995), "Economic Reform and the Process of Global Integration (with discussion)," *Brookings Papers on Economic Activity*, 1, 1-118.

Sachs, J and A. Warner, (1996), "Natural Resource Abundance and Economic Growth," *National Bureau of Economic Research Working Paper no. 5398*.

Sala-i-Martin, X., (1991), "Growth, Macroeconomics, and Development: Comments," in *Macroeconomics Annual 1991*, O. Blanchard and S. Fischer, eds., Cambridge: MIT Press.

Sala-i-Martin, X., (1996a), "The Classical Approach to Convergence Analysis," *Economic Journal*, 106, 1019-1036.

Sala-i-Martin, X., (1996b), "Regional Cohesion: Evidence and Theories of Regional Growth and Convergence," *European Economic Review*, 40, 1325-1352.

Sala-i-Martin, X., (1997a), "I Just Ran 4 Million Regressions," *National Bureau of Economic Research Working Paper no. 6252.*

Sala-i-Martin, X., (1997b), "I Just Ran 2 Million Regressions." *American Economic Review*, 87, 2, 178-83.

Shioji, E., (2001a), "Composition Effect of Migration and Regional Growth in Japan," *Journal of the Japanese and International Economies*, 15, 29-49.

Shioji, E., (2001b), "Public Capital and Economic Growth: A Convergence Approach," *Journal of Economic Growth*, 6, 3, 205-227.

Solow, R., (1956), "A Contribution to the Theory of Economic Growth," *Quarterly Journal of Economics*, 70, 1, 65-94.

Solow, R., (1994), "Perspectives on Growth Theory," *Journal of Economic Perspectives*, 8, 45-54.

Summers, R. and A. Heston, (1988), "A New Set of International Comparisons of Real Product and Price Levels Estimates for 130 Countries, 1950-1985," *Review of Income and Wealth*, 34, 1-25.

Summers, R. and A. Heston, (1991), "The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950-1988," *Quarterly Journal of Economics*, 106, 2, 327-368.

Swan, T., (1956), "Economic Growth and Capital Accumulation," *Economic Record*, 32, 334-361.

Swank, D., (1996), "Culture, Institutions, and Economic Growth," *American Journal of Political Science*, 40, 660-79.

Swartz, S. and R. Welsch, (1986), "Applications of Bounded-Influence and Diagnostic Methods in Energy Modeling," in *Model Reliability*, D. Belsley, and E. Kuh, eds., Cambridge: MIT Press.

Tan, C. M., (2004), "No One True Path to Development: Uncovering the Interplay Between Geography, Institutions, and Ethnic Fractionalization in Economic Development," mimeo, Tufts University.

Tavares, J. and R. Wacziarg, (2001), "How Democracy Affects Growth," *European Economic Review*, 45, 8, 1341-78.

Taylor, C., (1998), *Socrates*, New York: Oxford University Press.

Temple, J., (1998), "Robustness Tests of the Augmented Solow Model," *Journal of Applied Econometrics*, 13, 4, 361-75.

Temple, J., (1999), "The New Growth Evidence," *Journal of Economic Literature*, 37, 1, 112-56.

Temple, J., (2000a), "Inflation and Growth: Stories Short and Tall," *Journal of Economic Surveys*, 14, 4, 395-426.

Temple, J., (2000b), "Growth Regressions and What the Textbooks Don't Tell You," *Bulletin of Economic Research*, 52, 3, 181-205.

Temple, J., (2003), "The Long-run Implications of Growth Theories," *Journal of Economic Surveys*, 17, 3, 497-510.

Temple, J. and P. Johnson, (1998), "Social Capability and Economic Growth," *Quarterly Journal of Economics*, 113, 3, 965-90.

Toya, H., M. Skidmore, and R. Robertson, (2003), "Why are Estimates of Human Capital's Contribution to Growth So Small," mimeo, Nagoya City University.

Wacziarg, R., (2002), "Review of Easterly's The Elusive Quest for Growth," *Journal of Economic Literature*, 40, 3, 907-18.

Wacziarg, R. and K. Welch, (2003), "Trade Liberalization and Growth: New Evidence," *National Bureau of Economic Research Working Paper no. 10152.*

Warner, A., (1992), "Did the Debt Crisis Cause the Investment Crisis?," *Quarterly Journal of Economics*, 107, 4, 1161-86.

Welsch, H., (2003), "Corruption, Growth and the Environment: A Cross Country Analysis," mimeo, German Institute for Economic Research.

White, H., (1980), "A Heteroskedastic-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.

Whiteley, P. (2000), "Economic Growth and Social Capital," *Political Studies,* 48, 443-466.

Zak, P. and S. Knack, (2001), "Trust and Growth," *Economic Journal*, 111, 295-321.

Zietz, J., (2001), "Heteroskedasticity and Neglected Parameter Heterogeneity," *Oxford Bulletin of Economics and Statistics*, 63, 2, 263-73.

ACCOUNTING FOR GROWTH IN THE INFORMATION AGE

by

Dale W. Jorgenson

1. The Information Age.*

1.1. Introduction.

The resurgence of the American economy since 1995 has outrun all but the most optimistic expectations. Economic forecasting models have been seriously off track and growth projections have been revised repeatedly to reflect a more sanguine outlook[1]. It is not surprising that the unusual combination of more rapid growth and slower inflation touched off a strenuous debate about whether improvements in America's economic performance could be sustained.

The starting point for the economic debate is the thesis that the 1990's are a mirror image of the 1970's, when an unfavorable series of "supply shocks" led to stagflation -- slower growth and higher inflation[2]. In this view, the development of information technology (IT) is one of a series of positive, but

---

* Department of Economics, Harvard University, 122 Littauer Center, Cambridge, MA 02138-3001. The Program on Technology and Economic Policy at Harvard University provided financial support. The Economic and Social Research Institute of the Cabinet Office of the Government of Japan supported the research reported in Section 4 from its program for international collaboration through the Nomura Research Institute. I am greatly indebted to Jon Samuels for excellent research assistance, as well as useful comments. J. Steven Landefeld, Clinton McCully, and David Wasshausen of the Bureau of Economic Analysis provided valuable data on information technology in the U.S. Tom Hale, Mike Harper, Tom Nardone and Larry Rosenblum (BLS), Kurt Kunze (BEA), Eldon Ball (ERS), Mike Dove and Scott Segerman (DMDC) also provided data for the U.S. and helpful advice. I am grateful to John Baldwin and Tarek Harchaoui of Statistics Canada for data on Canada, Kazuyuki Motohashi and Koji Nomura for data on Japan, and Alessandra Colecchia, Marcel Timmer and Bart Van Ark for data on Europe. Colleagues far too numerous to mention have contributed useful suggestions. I am grateful to all of them but retain sole responsibility for any remaining deficiencies.

[1] See Congressional Budget Office (2000) on official forecasts and Economics and Statistics Administration (2000), p. 60, on private forecasts.

[2] Robert Gordon (1998, 2000); Barry Bosworth and Jack Triplett (2000).

temporary, shocks. The competing perspective is that IT has produced a fundamental change in the U.S. economy, leading to a *permanent* improvement in growth prospects[3]. The resolution of this debate in favor of a permanent improvement has been the "killer application" of a new framework for productivity measurement summarized in Paul Schreyer's (2001) OECD Manual, *Measuring Productivity*.

A consensus has emerged that the development and deployment of information technology is the foundation of the American growth resurgence. A mantra of the "new economy" -- *faster, better, cheaper* -- captures the speed of technological change and product improvement in semiconductors and the precipitous and continuing fall in semiconductor prices. The price decline has been transmitted to the prices of products that rely heavily on semiconductor technology, like computers and telecommunications equipment. This technology has also helped to reduce the cost of aircraft, automobiles, scientific instruments, and a host of other products.

Swiftly falling IT prices provide powerful economic incentives for the substitution of IT equipment for other forms of capital and for labor services. The rate of the IT price decline is a key component of the cost of capital, required for assessing the impacts of rapidly growing stocks of computers, communications equipment, and software. Constant quality price indexes are essential for identifying the change in price for a given level of performance. Accurate and timely computer prices have been part of the U.S. National Income and Product Accounts (NIPA) since 1985. Unfortunately, important information gaps remain, especially on trends in prices for closely related investments, such as software and communications equipment.

Capital input has been the most important source of U.S. economic growth throughout the postwar period. More rapid substitution toward information

[3] Alan Greenspan (2000).

technology has given much additional weight to components of capital input with higher marginal products. The vaulting contribution of capital input since 1995 has boosted growth by close to a percentage point. The contribution of investment in IT accounts for more than half of this increase. Computers have been the predominant impetus to faster growth, but communications equipment and software have made important contributions as well.

The accelerated information technology price decline signals faster productivity growth in IT-producing industries. In fact, these industries have been a rapidly rising source of aggregate productivity growth throughout the 1990's. The IT-producing industries generate less than five percent of gross domestic income, but have accounted for nearly half the surge in productivity growth since 1995. However, it is important to emphasize that faster productivity growth is not limited to these industries.

The dramatic effects of information technology on capital and labor markets have already generated a substantial and growing economic literature, but many important issues remain to be resolved. For capital markets the relationship between equity valuations and growth prospects merits much further study. For labor markets more research is needed on investment in information technology and substitution among different types of labor.

1.2. Faster, Better, Cheaper.

Modern information technology begins with the invention of the *transistor*, a semiconductor device that acts as an electrical switch and encodes information in binary form. A binary digit or *bit* takes the values zero and one, corresponding to the off and on positions of a switch. The first transistor, made of the semiconductor germanium, was constructed at Bell Labs in 1947 and

won the Nobel Prize in Physics in 1956 for the inventors -- John Bardeen, Walter Brattain, and William Shockley[4].

The next major milestone in information technology was the co-invention of the *integrated circuit* by Jack Kilby of Texas Instruments in 1958 and Robert Noyce of Fairchild Semiconductor in 1959. An integrated circuit consists of many, even millions, of transistors that store and manipulate data in binary form. Integrated circuits were originally developed for data storage and retrieval and semiconductor storage devices became known as *memory chips*[5].

The first patent for the integrated circuit was granted to Noyce. This resulted in a decade of litigation over the intellectual property rights. The litigation and its outcome demonstrate the critical importance of intellectual property in the development of information technology. Kilby was awarded the Nobel Prize in Physics in 2000 for discovery of the integrated circuit; regrettably, Noyce died in 1990[6].

1.2.1. Moore's Law.

In 1965 Gordon Moore, then Research Director at Fairchild Semiconductor, made a prescient observation, later known as *Moore's Law*[7]. Plotting data on memory chips, he observed that each new chip contained roughly twice as many transistors as the previous chip and was released within 18-24 months of its predecessor. This implied exponential growth of chip capacity at 35-45 percent per year! Moore's prediction, made in the infancy of the semiconductor industry, has tracked chip capacity for thirty-five years. He recently extrapolated this trend for at least another decade[8].

---

[4] On Bardeen, Brattain, and Shockley, see: http://www.nobel.se/physics/laureates/1956/.

[5] Charles Petzold (2000) provides a general reference on computers and software.

[6] On Kilby, see: http://www.nobel.se/physics/laureates/2000/. On Noyce, see: Tom Wolfe (2000), pp. 17-65.

[7] Moore (1965). Vernon Ruttan (2001), pp.316-367, provides a general reference on the economics of semiconductors and computers. On semiconductor technology, see: http://euler.berkeley.edu/~esrc/csm.

[8] Moore (1997).

In 1968 Moore and Noyce founded Intel Corporation to speed the commercialization of memory chips[9]. Integrated circuits gave rise to *microprocessors* with functions that can be programmed by software, known as *logic chips*. Intel's first general purpose microprocessor was developed for a calculator produced by Busicom, a Japanese firm. Intel retained the intellectual property rights and released the device commercially in 1971.

The rapidly rising trends in the capacity of microprocessors and storage devices illustrate the exponential growth predicted by Moore's Law. The first logic chip in 1971 had 2,300 transistors, while the Pentium 4 released on November 20, 2000, had 42 million! Over this twenty-nine year period the number of transistors increased by thirty-four percent per year. The rate of productivity growth for the U.S. economy during this period was slower by two orders of magnitude.

1.2.2. Semiconductor Prices.

Moore's Law captures the fact that successive generations of semiconductors are *faster* and *better*. The economics of semiconductors begins with the closely related observation that semiconductors have become *cheaper* at a truly staggering rate! Figure 1.1 gives semiconductor price indexes constructed by Bruce Grimm (1998) of the Bureau of Economic Analysis (BEA) and employed in the U.S. National Income and Product Accounts since 1996. These are divided between memory chips and logic chips. The underlying detail includes seven types of memory chips and two types of logic chips.

Between 1974 and 1996 prices of memory chips *decreased* by a factor of 27,270 times or at 40.9 percent per year, while the implicit deflator for the gross domestic product (GDP) *increased* by almost 2.7 times or 4.6 percent per year! Prices of logic chips, available for the shorter period 1985 to 1996, *decreased* by a factor of 1,938 or 54.1 percent per year, while the GDP deflator

[9] Moore (1996).

increased by 1.3 times or 2.6 percent per year! Semiconductor price declines closely parallel Moore's Law on the growth of chip capacity, setting semiconductors apart from other products.

Figure 1.1 also reveals a sharp acceleration in the decline of semiconductor prices in 1994 and 1995. The microprocessor price decline leapt to more than ninety percent per year as the semiconductor industry shifted from a three-year product cycle to a greatly accelerated two-year cycle. This is reflected in the *2000 Update* of the International Technology Road Map for Semiconductors[10], prepared by a consortium of industry associations. Ana Aizcorbe, Stephen Oliner, and Daniel Sichel (2003) have identified and analyzed break points in prices of microprocessors and storage devices.

1.2.3. Constant Quality Price Indexes.

The behavior of semiconductor prices is a severe test for the methods used in the official price statistics. The challenge is to separate observed price changes between changes in semiconductor performance and changes in price that hold performance constant. Achieving this objective has required a detailed understanding of the technology, the development of sophisticated measurement techniques, and the introduction of novel methods for assembling the requisite information.

Ellen Dulberger (1993) introduced a "matched model" index for semiconductor prices. A matched model index combines price relatives for products with the same performance at different points of time. Dulberger presented constant quality price indexes based on index number formulas, including the *Fisher* (1922) *ideal index* used in the in the U.S. national accounts[11]. The Fisher index is the geometric average of the familiar Laspeyres and Paasche indexes.

---

[10] On International Technology Roadmap for Semiconductors (2000), see: http://public.itrs.net/.

[11] See Steven Landefeld and Robert Parker (1997).

Erwin Diewert (1976) defined a *superlative* index number as an index that *exactly* replicates a *flexible* representation of the underlying technology (or preferences). A flexible representation provides a second-order approximation to an arbitrary technology (or preference system). A.A. Konus and S. S. Byushgens (1926) first showed that the Fisher ideal index is superlative in this sense. Laspeyres and Paasche indexes are not superlative and fail to capture substitutions among products in response to price changes accurately.

Grimm (1998) combined matched model techniques with hedonic methods, based on an econometric model of semiconductor prices at different points of time. A hedonic model gives the price of a semiconductor product as a function of the characteristics that determine performance, such as speed of processing and storage capacity. A constant quality price index isolates the price change by holding these characteristics of semiconductors fixed.[12]

Beginning in 1997, the Bureau of Labor Statistics (BLS) incorporated a matched model price index for semiconductors into the Producer Price Index (PPI) and since then the national accounts have relied on data from the PPI. Reflecting long-standing BLS policy, historical data were not revised backward. Semiconductor prices reported in the PPI prior to 1997 do not hold quality constant, failing to capture the rapid semiconductor price decline and the acceleration in 1995.

1.2.4. Computers.

The introduction of the Personal Computer (PC) by IBM in 1981 was a watershed event in the deployment of information technology. The sale of Intel's 8086-8088 microprocessor to IBM in 1978 for incorporation into the PC was a major business breakthrough for Intel[13]. In 1981 IBM licensed the MS-DOS operating system from the Microsoft Corporation, founded by Bill Gates and Paul

[12]Triplett (2003) has drafted a manual for the OECD on constructing constant quality price indexes for information technology and communications equipment and software.
[13]See Moore (1996).

Allen in 1975. The PC established an Intel/Microsoft relationship that has continued up to the present. In 1985 Microsoft released the first version of Windows, its signature operating system for the PC, giving rise to the Wintel (Windows-Intel) nomenclature for this ongoing collaboration.

Mainframe computers, as well as PC's, have come to rely heavily on logic chips for central processing and memory chips for main memory. However, semiconductors account for less than half of computer costs and computer prices have fallen much less rapidly than semiconductor prices. Precise measures of computer prices that hold product quality constant were introduced into the NIPA in 1985 and the PPI during the 1990's. The national accounts now rely on PPI data, but historical data on computers from the PPI, like the PPI data on semiconductors, do not hold quality constant.

Gregory Chow (1967) pioneered the use of hedonic techniques for constructing a constant quality index of computer prices in research conducted at IBM. Chow documented price declines at more than twenty percent per year during 1960-1965, providing an initial glimpse of the remarkable behavior of computer prices. In 1985 the Bureau of Economic Analysis incorporated constant quality price indexes for computers and peripheral equipment constructed by IBM into the NIPA. Triplett's (1986) discussion of the economic interpretation of these indexes brought the rapid decline of computer prices to the attention of a very broad audience.

The BEA-IBM constant quality price index for computers provoked a heated exchange between BEA and Edward Denison (1989), one of the founders of national accounting methodology in the 1950's and head of the national accounts at BEA from 1979 to 1982. Denison sharply attacked the BEA-IBM methodology and argued vigorously against the introduction of constant quality price indexes into the national accounts[14]. Allan Young (1989), then Director of BEA, reiterated BEA's

[14] Denison cited his 1957 paper, "Theoretical Aspects of Quality Change, Capital

rationale for introducing constant quality price indexes.

Dulberger (1989) presented a more detailed report on her research on the prices of computer processors for the BEA-IBM project. Speed of processing and main memory played central roles in her model. Triplett (1989, 2003) has provided exhaustive surveys of research on hedonic price indexes for computers. Gordon (1989, 1990) gave an alternative model of computer prices and identified computers and communications equipment, along with commercial aircraft, as assets with the highest rates of price decline.

Figure 1.2 gives BEA's constant quality index of prices of computers and peripheral equipment and its components, including mainframes, PC's, storage devices, other peripheral equipment, and terminals. The decline in computer prices follows the behavior of semiconductor prices presented in Figure 1.1, but in much attenuated form. The 1995 acceleration in the computer price decline parallels the acceleration in the semiconductor price decline that resulted from the changeover from a three-year product cycle to a two-year cycle in 1995.

1.2.5. Communications Equipment and Software.

Communications technology is crucial for the rapid development and diffusion of the Internet, perhaps the most striking manifestation of information technology in the American economy[15]. Kenneth Flamm (1989) was the first to compare the behavior of computer prices and the prices of communications equipment. He concluded that the communications equipment prices fell only a little more slowly than computer prices. Gordon (1990) compared Flamm's results with the official price indexes, revealing substantial bias in the official indexes.

Consumption, and Net Capital Formation," as the definitive statement of the traditional BEA position.

[15] General references on the economics of the Internet are Soon-Yong Choi and Andrew Whinston (2000) and Robert Hall (2002). On Internet indicators see: http://www.internetindicators.com/.

Communications equipment is an important market for semiconductors, but constant quality price indexes cover only a portion of this equipment. Switching and terminal equipment rely heavily on semiconductor technology, so that product development reflects improvements in semiconductors. Grimm's (1997) constant quality price index for digital telephone switching equipment, given in Figure 3.3, was incorporated into the national accounts in 1996. The output of communications services in the NIPA also incorporates a constant quality price index for cellular phones.

Much communications investment takes the form of the transmission gear, connecting data, voice, and video terminals to switching equipment. Technologies such as fiber optics, microwave broadcasting, and communications satellites have progressed at rates that outrun even the dramatic pace of semiconductor development. An example is dense wavelength division multiplexing (DWDM), a technology that sends multiple signals over an optical fiber simultaneously. Installation of DWDM equipment, beginning in 1997, has doubled the transmission capacity of fiber optic cables every 6-12 months[16].

Mark Doms (2004) has provided comprehensive price indexes for terminals, switching gear, and transmission equipment. These have been incorporated into the Federal Reserve's Index of Industrial Production, as described by Carol Corrado (2003), but are not yet included in the U.S. National Income and Product Accounts. The analysis of the impact of information technology on the U.S. economy described below is based on the national accounts and remains incomplete.

Both software and hardware are essential for information technology and this is reflected in the large volume of software expenditures. The eleventh comprehensive revision of the national accounts, released by BEA on October 27,

[16] Rick Rashad (2000) characterizes this as the "demise" of Moore's Law. Jeff Hecht (1999) describes DWDM technology and provides a general reference on fiber optics.

1999, re-classified computer software as investment[17]. Before this important advance, business expenditures on software were treated as current outlays, while personal and government expenditures were treated as purchases of nondurable goods. Software investment is growing rapidly and is now much more important than investment in computer hardware.

Parker and Grimm (2000) describe the new estimates of investment in software. BEA distinguishes among three types of software -- prepackaged, custom, and own-account software. Prepackaged software is sold or licensed in standardized form and is delivered in packages or electronic files downloaded from the Internet. Custom software is tailored to the specific application of the user and is delivered along with analysis, design, and programming services required for customization. Own-account software consists of software created for a specific application. However, only price indexes for prepackaged software hold performance constant.

Parker and Grimm (2000) present a constant quality price index for prepackaged software, given in Figure 3.3. This combines a hedonic model of prices for business applications software and a matched model index for spreadsheet and word processing programs developed by Oliner and Sichel (1994). Prepackaged software prices decline at more than ten percent per year over the period 1962-1998. Since 1998 the BEA has relied on a matched model price index for all prepackaged software from the PPI; prior to 1998 the PPI data do not hold quality constant.

BEA's prices for own-account and custom software are based on programmer wage rates. This implicitly assumes no change in the productivity of computer programmers, even with growing investment in hardware and software to support the creation of new software. Custom and own-account software prices are a

[17] Brent Moulton (2000) describes the 11th comprehensive revision of NIPA and the 1999 update.

weighted average of prepackaged software prices and programmer wage rates with arbitrary weights of 75 percent for programmer wage rates and 25 percent for prepackaged software. These price indexes do not hold the software performance constant and present a distorted picture of software prices, as well as software output and investment.

1.2.6. Research Opportunities.

The official price indexes for computers and semiconductors provide the paradigm for economic measurement. These indexes capture the steady decline in IT prices and the recent acceleration in this decline. The official price indexes for central office switching equipment and prepackaged software also hold quality constant. BEA and BLS, the leading statistical agencies in price research, have carried out much of the best work in this area. However, a critical role has been played by price research at IBM, long the dominant firm in information technology[18].

It is important to emphasize that information technology is not limited to applications of semiconductors. Switching and terminal equipment for voice, data, and video communications have come to rely on semiconductor technology and the empirical evidence on prices of this equipment reflects this fact. Transmission gear employs technologies with rates of progress that far outstrip those of semiconductors. This important gap in our official price statistics has been filled by constant quality price indexes for all types of communications equipment constructed by Doms (2004), but these indexes have not been incorporated into the national accounts.

Investment in software is more important than investment in hardware. This was essentially invisible until BEA introduced new measures of prepackaged, custom, and own-account software investment into the national accounts in 1999. This is a crucial step in understanding the role of information technology in

[18] See Alfred Chandler (2000), Table 1.1, p. 26.

the American economy. Unfortunately, software prices are a statistical blind spot with only prices of prepackaged software adequately represented in the official system of price statistics. The daunting challenge that lies ahead is to construct constant quality price indexes for custom and own-account software.

1.3. Impact of Information Technology.

In Section 2 I consider the "killer application" of the new framework for productivity measurement – the impact of information technology (IT) on economic growth. Despite differences in methodology and data sources, a consensus has emerged that the remarkable behavior of IT prices provides the key to the surge in U.S. economic growth after 1995. The relentless decline in the prices of information technology equipment and software has steadily enhanced the role of IT investment. Productivity growth in IT-producing industries has risen in importance and a productivity revival is underway in the rest of the economy.

A substantial acceleration in the IT price decline occurred in 1995, triggered by a much sharper acceleration in the price decline of semiconductors, the key component of modern information technology. Although the decline in semiconductor prices has been projected to continue for at least another decade, the recent acceleration may be temporary. This can be traced to a shift in the product cycle for semiconductors from three years to two years as a consequence of intensifying competition in markets for semiconductor products.

In Section 3 I show that the surge of IT investment in the United States after 1995 has counterparts in all other industrialized countries. It is essential to use comparable data and methodology in order to provide rigorous international comparisons. A crucial role is played by measurements of IT prices. The U.S. national accounts have incorporated measures of IT prices that hold performance constant since 1985. Schreyer (2000) has extended these

measures to other industrialized countries by constructing "internationally harmonized prices".[19]

I show that the acceleration in the IT price decline in 1995 triggered a burst of IT investment in all of the G7 nations – Canada, France, Germany, Italy, Japan, the U.K., as well as the U.S. These countries also experienced a rise in productivity growth in the IT-producing industries. However, differences in the relative importance of these industries have generated wide disparities in the impact of IT on economic growth. The role of the IT-producing industries is greatest in the U.S., which leads the G7 in output per capita. Section 4 concludes.

2. Aggregate Growth Accounting.

2.1. The Role of Information Technology.

At the aggregate level IT is identified with the outputs of computers, communications equipment, and software. These products appear in the GDP as investments by businesses, households, and governments along with net exports to the rest of the world. The GDP also includes the services of IT products consumed by households and governments. A methodology for analyzing economic growth must capture the substitution of IT outputs for other outputs of goods and services.

While semiconductor technology is the driving force behind the spread of IT, the impact of the relentless decline in semiconductor prices is transmitted through falling IT prices. Only net exports of semiconductors, defined as the difference between U.S. exports to the rest of the world and U.S. imports appear in the GDP. Sales of semiconductors to domestic manufacturers of IT products are precisely offset by purchases of semiconductors and are excluded from the GDP.

[19] The measurement gap in IT prices between the U.S. and other OECD countries was first identified by Andrew Wyckoff (1995).

Constant quality price indexes, like those reviewed in the previous section, are a key component of the methodology for analyzing the American growth resurgence. Computer prices were incorporated into the NIPA in 1985 and are now part of the PPI as well. Much more recently, semiconductor prices have been included in the NIPA and the PPI. The official price indexes for communications equipment do not yet reflect the important work of Doms (2004). Unfortunately, evidence on the price of software is seriously incomplete, so that the official price indexes are seriously misleading.

2.1.1. Output.

The output data in Table 2.1 are based on the most recent benchmark revision of the national accounts through 2002[20]. The output concept is similar, but not identical, to the concept of gross domestic product used by the BEA. Both measures include final outputs purchased by businesses, governments, households, and the rest of the world. Unlike the BEA concept, the output measure in Table 2.1 also includes imputations for the service flows from durable goods, including IT products, employed in the household and government sectors.

The imputations for services of IT equipment are based on the cost of capital for IT described in more detail below. The cost of capital is multiplied by the nominal value of IT capital stock to obtain the imputed service flow from IT products. In the business sector this accrues as capital income to the firms that employ these products as inputs. In the household and government sectors the flow of capital income must be imputed. This same type of imputation is used for housing in the NIPA. The rental value of renter-occupied housing accrues to real estate firms as capital income, while the rental value of owner-occupied housing is imputed to households.

[20] See Jorgenson and Stiroh (2000b), Appendix A, for details on the estimates of output.

Current dollar GDP in Table 2.1 is $11.3 trillions in 2002, including imputations, and real output growth averaged 3.46 percent for the period 1948-2002. These magnitudes can be compared to the current dollar value of $10.5 trillions in 2002 and the average real growth rate of 3.36 percent for period 1948-2002 for the official GDP. Table 2.1 presents the current dollar value and price indexes of the GDP and IT output. This includes outputs of investment goods in the form of computers, software, communications equipment, and non-IT investment goods. It also includes outputs of non-IT consumption goods and services as well as imputed IT capital service flows from households and governments.

The most striking feature of the data in Table 2.1 is the rapid price decline for computer investment, 15.8 percent per year from 1959 to 1995. Since 1995 this decline has increased to 33.1 percent per year. By contrast the relative price of software has been flat for much of the period and began to fall only in the 1980's. The price of communications equipment behaves similarly to the software price, while the consumption of capital services from computers and software by households and governments shows price declines similar to computer investment.

The top panel of Table 2.2 summarizes the growth rates of prices and quantities for major output categories for 1989-95 and 1995-2002. Business investments in computers, software, and communications equipment are the largest categories of IT spending. Households and governments have also spent sizable amounts on computers, software, communications equipment and the services of information technology. Figure 2.1 shows that the share of software output in the GDP is largest, followed by the shares of computers and communications equipment.

2.1.2. Capital Services.

This section presents capital estimates for the U.S. economy for the period 1948 to 2002[21]. These begin with BEA investment data; the perpetual inventory method generates estimates of capital stocks and these are aggregated, using service prices as weights. This approach, originated by Jorgenson and Zvi Griliches (1967), is based on the identification of service prices with marginal products of different types of capital. The service price estimates incorporate the cost of capital[22].

The cost of capital is an annualization factor that transforms the price of an asset into the price of the corresponding capital input. This includes the nominal rate of return, the rate of depreciation, and the rate of capital loss due to declining prices. The cost of capital is an essential concept for the economics of information technology[23], due to the astonishing decline of IT prices given in Table 2.1.

The cost of capital is important in many areas of economics, especially in modeling producer behavior, productivity measurement, and the economics of taxation[24]. Many of the important issues in measuring the cost of capital have been debated for decades. The first of these is incorporation of the rate of decline of asset prices into the cost of capital. The assumption of perfect foresight or rational expectations quickly emerged as the most appropriate formulation and has been used in almost all applications of the cost of capital[25].

---

[21] See Jorgenson and Stiroh (2000b), Appendix B, for details on the estimates of capital input.

[22] Jorgenson and Kun-Young Yun (2001) present the model of capital input used in the estimates presented in this section. BLS (1983) describes the version of this model employed in the official productivity statistics. For a recent updates, see the BLS multifactor productivity website: http://www.bls.gov/mfp/home.htm. Charles Hulten (2001) surveys the literature.

[23] Jorgenson and Stiroh (1995), pp. 300-303.

[24] Lawrence Lau (2000) surveys applications of the cost of capital.

[25] See, for example, Jorgenson, Gollop, and Fraumeni (1987), pp. 40-9, and Jorgenson and Griliches (1967).

The second empirical issue is the measurement of economic depreciation. The stability of patterns of depreciation in the face of changes in tax policy and price shocks has been carefully documented. The depreciation rates presented by Jorgenson and Stiroh (2000b) summarize a large body of empirical research on the behavior of asset prices[26]. A third empirical issue is the description of the tax structure for capital income. This depends on the tax laws prevailing at each point of time. The resolution of these issues has cleared the way for detailed measurements of the cost of capital for all assets that appear in the national accounts, including information technology equipment and software[27].

The definition of capital includes all tangible assets in the U.S. economy, equipment and structures, as well as consumers' and government durables, land, and inventories. The capital service flows from durable goods employed by households and governments enter measures of both output and input. A steadily rising proportion of these service flows are associated with investments in IT. Investments in IT by business, household, and government sectors must be included in the GDP, along with household and government IT capital services, in order to capture the full impact of IT on the U.S. economy.

Table 2.3 gives capital stocks from 1948 to 2002, as well as price indexes for total domestic tangible assets and IT assets -- computers, software, and communications equipment. The estimate of domestic tangible capital stock in Table 2.3 is $45.9 trillions in 2002, considerably greater than the estimate by BEA. The most important differences reflect the inclusion of inventories and land in Table 2.3.

---

[26] Jorgenson and Stiroh (2000b), Table B4, pp. 196-7 give the depreciation rates employed in this section. Fraumeni (1997) describes depreciation rates used in the NIPA. Jorgenson (1996) surveys empirical studies of depreciation.

[27] See Jorgenson and Yun (2001) for details on the U.S. tax structure for capital income. Diewert and Denis Lawrence (2000) survey measures of the price and quantity of capital input.

Business IT investments, as well as purchases of computers, software, and communications equipment by households and governments, have grown spectacularly in recent years, but remain relatively small. The stocks of all IT assets combined account for only 3.79 percent of domestic tangible capital stock in 2002. Table 2.4 presents estimates of the flow of capital services and corresponding price indexes for 1948-2002.

The difference between growth in capital services and capital stock is the improvement in capital quality. This represents the substitution towards assets with higher marginal products. The shift toward IT increases the quality of capital, since computers, software, and communications equipment have relatively high marginal products. Capital stock estimates fail to account for this increase in quality and substantially underestimate the impact of IT investment on growth.

The growth of capital quality is slightly more than twenty percent of capital input growth for the period 1948-2002. However, improvements in capital quality have increased steadily in relative importance. These improvements jumped to 46.1 percent of total growth in capital input during the period 1995-2002, reflecting very rapid restructuring of capital to take advantage of the sharp acceleration in the IT price decline. Capital stock has become progressively less accurate as a measure of capital input and is now seriously deficient.

Figure 2.2 gives the IT capital service flows as a share of gross domestic income. The second panel of Table 2.2 summarizes the growth rates of prices and quantities of capital inputs for 1989-1995 and 1995-2002. Growth of IT capital services jumps from 12.39 percent per year in 1989-1995 to 18.11 percent in 1995-2002, while growth of non-IT capital services increases from 1.94 percent to 3.07 percent. This reverses the trend toward slower capital growth through 1995.

2.1.3. Labor Services.

This section presents estimates of labor input for the U.S. economy from 1948 to 2002. These incorporate individual data from the Censuses of Population for 1970, 1980, and 1990, as well as the annual Current Population Surveys. Constant quality indexes for the price and quantity of labor input account for the heterogeneity of the workforce across sex, employment class, age, and education levels. This follows the approach of Jorgenson, Gollop, and Fraumeni (1987)[28].

The distinction between labor input and labor hours is analogous to the distinction between capital services and capital stock. The growth in labor quality is the difference between the growth in labor input and hours worked. Labor quality reflects the substitution of workers with high marginal products for those with low marginal products. Table 2.5 presents estimates of labor input, hours worked, and labor quality.

The value of labor expenditures in Table 2.5 is $6.6 trillions in 2002, 58.3 percent of the value of output. This share accurately reflects the concept of gross domestic income, including imputations for the value of capital services in household and government sectors. As shown in Table 2.7, the growth rate of labor input decelerated to 1.50 percent for 1995-2002 from 1.64 percent for 1989-1995. Growth in hours worked rose from 1.02 percent for 1989-1995 to 1.16 percent for 1995-2002 as labor force participation increased and unemployment rates declined.

The growth of labor quality has declined considerably since 1995, dropping from 0.61 percent for 1989-1995 to 0.33 percent for 1995-2002. This slowdown captures well-known demographic trends in the composition of the work force, as well as exhaustion of the pool of available workers. Growth in hours worked does

[28]See Jorgenson and Stiroh (2000b), Appendix C, for details on the estimates of labor input. Gollop (2000) discusses the measurement of labor quality.

not capture these changes in labor quality growth and is a seriously misleading measure of labor input.

2.2. The American Growth Resurgence.

The American economy has undergone a remarkable resurgence since the mid-1990's with accelerating growth in output, labor productivity, and total factor productivity. The purpose of this section is to quantify the sources of growth for 1948-2002 and various sub-periods. An important objective is to account for the sharp acceleration in the growth rate since 1995 and, in particular, to document the role of information technology.

The appropriate framework for analyzing the impact of information technology is the production possibility frontier, giving outputs of IT investment goods as well as inputs of IT capital services. An important advantage of this framework is that prices of IT outputs and inputs are linked through the price of IT capital services. This framework successfully captures the substitutions among outputs and inputs in response to the rapid deployment of IT. It also encompasses costs of adjustment, while allowing financial markets to be modeled independently.

As a consequence of the swift advance of information technology, a number of the most familiar concepts in growth economics have been superseded. The aggregate production function heads this list. Capital stock as a measure of capital input is no longer adequate to capture the rising importance of IT. This completely obscures the restructuring of capital input that is such an important wellspring of the growth resurgence. Finally, hours worked must be replaced as a measure of labor input.

2.2.1. Production Possibility Frontier.

The production possibility frontier describes efficient combinations of outputs and inputs for the economy as a whole. Aggregate output Y consists of outputs of investment goods and consumption goods. These outputs are produced from aggregate input X, consisting of capital services and labor services.

Productivity is a "Hicks-neutral" augmentation of aggregate input. The

$$Y(I_n, I_c, I_s, I_t, C_n, C_c) = A \cdot X(K_n, K_c, K_s, K_t, L),$$

production possibility frontier takes the form:

where the outputs include non-IT investment goods I_n and investments in computers I_c, software I_s, and communications equipment I_t, as well as non-IT consumption goods and services C_n and IT capital services to households and governments C_c. Inputs include non-IT capital services K_n and the services of computers K_c, software K_s, and telecommunications equipment K_t , as well as labor input L.[29] Productivity is denoted by A.

The most important advantage of the production possibility frontier is the explicit role that it provides for constant quality prices of IT products. These are used as deflators for nominal expenditures on IT investments to obtain the quantities of IT outputs. Investments in IT are cumulated into stocks of IT capital. The flow of IT capital services is an aggregate of these stocks with service prices as weights. Similarly, constant quality prices of IT capital services are used in deflating the nominal values of consumption of these services.

Another important advantage of the production possibility frontier is the incorporation of costs of adjustment. For example, an increase in the output of IT investment goods requires foregoing part of the output of consumption goods and non-IT investment goods, so that adjusting the rate of investment in IT is costly. However, costs of adjustment are external to the producing unit and are fully reflected in IT prices. These prices incorporate forward-looking expectations of the future prices of IT capital services.

The aggregate production function employed, for example, by Kuznets (1971) and Solow (1957, 1960, 1970) and, more recently, by Jeremy Greenwood, Zvi

[29] Services of durable goods to governments and households are included in both inputs and outputs.

Hercowitz, and Per Krusell (1997, 2000), Hercowitz (1998), and Arnold Harberger (1998) is a competing methodology. The production function gives a single output as a function of capital and labor inputs. There is no role for separate prices of investment and consumption goods and, hence, no place for constant quality IT price indexes for outputs of IT investment goods.

Another limitation of the aggregate production function is that it fails to incorporate costs of adjustment. Robert Lucas (1967) presented a production model with internal costs of adjustment. Fumio Hayashi (2000) shows how to identify these adjustment costs from Tobin's (1969) Q-ratio, the ratio of the stock market value of the producing unit to the market value of the unit's assets. Implementation of this approach requires simultaneous modeling of production and asset valuation. If costs of adjustment are external, as in the production possibility frontier, asset valuation can be modeled separately from production[30].

2.2.2. Sources of Growth.

Under the assumption that product and factor markets are competitive producer equilibrium implies that the share-weighted growth of outputs is the sum of the share-weighted growth of inputs and growth in total factor productivity:

$$\overline{w}_{I,n}\Delta\ln I_n + \overline{w}_{I,c}\Delta I_c + \overline{w}_{I,s}\Delta I_s + \overline{w}_{I,t}\Delta I_t + \overline{w}_{C,n}C_n + \overline{w}_{C,c}\Delta\ln C_c =$$
$$\overline{v}_{K,n}\Delta\ln K_n + \overline{v}_{K,c}\Delta\ln K_c + \overline{v}_{K,s}\Delta\ln K_s + \overline{v}_{K,t}\Delta\ln K_t + \overline{v}_L\Delta\ln L + \Delta\ln A$$

where $\overline{w}$ and $\overline{v}$ denote average value shares. The shares of outputs and inputs add to one under the additional assumption of constant returns,

$$\overline{w}_{I,n} + \overline{w}_{I,c} + \overline{w}_{I,s} + \overline{w}_{I,t} + \overline{w}_{C,n} + \overline{w}_{C,c} = \overline{v}_{K,n} + \overline{v}_{K,c} + \overline{v}_{K,s} + \overline{v}_{K,t} + \overline{v}_L = 1 .$$

The growth rate of output is a weighted average of growth rates of investment and consumption goods outputs. The contribution of each output is its

[30] See, for example, John Campbell and Robert Shiller (1998).

weighted growth rate. Similarly, the growth rate of input is a weighted average of growth rates of capital and labor services and the contribution of each input is its weighted growth rate. The contribution of productivity, the growth rate of the augmentation factor A, is the difference between growth rates of output and input.

Table 2.6 presents results of a growth accounting decomposition for the period 1948-2002 and various sub-periods, following Jorgenson and Stiroh (1999, 2000b). Economic growth is broken down by output and input categories, quantifying the contribution of information technology to investment and consumption outputs, as well as capital inputs. These estimates identify computers, software, and communications equipment as distinct types of information technology.

The results can also be presented in terms of average labor productivity (ALP), defined as $y = Y / H$, the ratio of output Y to hours worked H, and $k = K / H$ is the ratio of capital services K to hours worked:

$$\Delta \ln y = \bar{v}_K \Delta \ln k + \bar{v}_L \left(\Delta \ln L - \Delta \ln H \right) + \Delta \ln A .$$

This equation allocates ALP growth among three sources. The first is capital deepening, the growth in capital input per hour worked, and reflects the capital-labor substitution. The second is improvement in labor quality and captures the rising proportion of hours by workers with higher marginal products. The third is total factor productivity growth, which contributes point-for-point to ALP growth.

2.2.3. Contributions of IT Investment.

Figure 2.2 depicts the rapid increase in the importance of IT services, reflecting the accelerating pace of IT price declines. In 1995-2002 the capital service price for computers fell 25.92 percent per year, compared to an increase of 32.09 percent in capital input from computers. While the value of computer

services grew, the current dollar value was only 1.13 percent of gross domestic income in 2002.

The rapid accumulation of software appears to have different sources. The price of software services has fallen only 1.48 percent per year for 1995-2002. Nonetheless, firms have been accumulating software very rapidly, with real capital services growing 14.02 percent per year. A possible explanation is that firms respond to computer price declines by investing in complementary inputs like software. However, a more plausible explanation is that the price indexes used to deflate software investment fail to hold quality constant. This leads to an overstatement of inflation and an understatement of growth.

Although the price decline for communications equipment during the period 1995-2002 is greater than that of software, investment in this equipment is more in line with prices. However, prices of communications equipment also fail to hold quality constant. The technology of switching equipment, for example, is similar to that of computers; investment in this category is deflated by a constant-quality price index developed by BEA. Conventional price deflators are employed for transmission gear, such as fiber-optic cables. This leads to an underestimate of the growth rates of investment, capital stock, capital services, and the GDP, as well as an overestimate of the rate of inflation.

Figures 2.3 and 2.4 highlight the rising contributions IT outputs to U.S. economic growth. Figure 2.3 shows the breakdown between IT and non-IT outputs for sub-periods from 1948 to 2002, while Figure 2.4 decomposes the contribution of IT into its components. Although the importance of IT has steadily increased, Figure 2.3 shows that the recent investment and consumption surge nearly doubled the output contribution of IT. Figure 2.4 shows that computer investment is the largest single IT contributor after 1995, but that investments in software and communications equipment are becoming increasingly important.

Figures 2.5 and 2.6 present a similar decomposition of IT inputs into production. The contribution of these inputs is rising even more dramatically.

Figure 2.5 shows that the contribution of IT now accounts for more than 48.0 percent of the total contribution of capital input. Figure 2.6 reveals that computer hardware is the largest component of IT, reflecting the growing share and accelerating growth rate of computer investment in the late 1990's.

Private business investment predominates in the output of IT, as shown by Jorgenson and Stiroh (2000b) and Oliner and Sichel (2000)[31]. Household purchases of IT equipment and services are next in importance. Government purchases of IT equipment and services, as well as net exports of IT products, must be included in order to provide a complete picture. Firms, consumers, governments, and purchasers of U.S. exports are responding to relative price changes, increasing the contributions of computers, software, and communications equipment.

Table 2.2 shows that the price of computer investment fell by 30.99 percent per year, the price of software fell by 1.31 percent, the price of communications equipment dropped by 4.16 percent, and the price of IT services fell by 13.91 percent during the period 1995-2002, while non-IT investment goods prices rose 0.38 percent. In response to these price changes, firms, households, and governments have accumulated computers, software, and communications equipment much more rapidly than other forms of capital.

2.2.4. Productivity.

The price or "dual" approach to productivity measurement employed by Triplett (1996) makes it possible to identify the role of IT production as a source of productivity growth at the industry level[32]. The rate of productivity growth is measured as the decline in the price of output, plus a weighted average of the growth rates of input prices with value shares of the inputs as weights. For the computer industry this expression is dominated by two terms:

---

[31] Bosworth and Triplett (2000) and Baily (2002) compare the results of Jorgenson and Stiroh with those of Oliner and Sichel, who incorporate data from the BLS measures of multifactor productivity.

[32] The dual approach is presented by Jorgenson, Gollop, and Fraumeni (1987), pp. 53-63.

the decline in the price of computers and the contribution of the price of semiconductors. For the semiconductor industry the expression is dominated by the decline in the price of semiconductors[33].

Jorgenson, Gollop, and Fraumeni (1987) have employed Domar's (1961) model to trace aggregate productivity growth to its sources at the level of individual industries[34]. More recently, Harberger (1998), William Gullickson and Michael Harper (1999), and Jorgenson and Stiroh (2000a, 2000b) have used the model for similar purposes. Productivity growth for each industry is weighted by the ratio of the gross output of the industry to GDP to estimate the industry contribution to aggregate productivity growth.

If semiconductor output were only used to produce computers, then its contribution to computer industry productivity growth, weighted by computer industry output, would precisely offset its independent contribution to the growth of aggregate productivity. This is the ratio of the value of semiconductor output to GDP, multiplied by the rate of semiconductor price decline. In fact, semiconductors are used to produce telecommunications equipment and many other products. However, the value of semiconductor output is dominated by inputs into IT production.

The Domar aggregation formula can be approximated by expressing the declines in prices of computers, communications equipment, and software relative to the price of gross domestic income, an aggregate of the prices of capital and labor services. The rates of relative IT price decline are weighted by ratios of the outputs of IT products to the GDP. Table 2.8 reports details of this decomposition of productivity for 1989-1995 and 1995-2002; the IT and non-IT contributions are presented in Figure 2.7. The IT products contribute 0.47

---

[33]Models of the relationships between computer and semiconductor industries presented by Dulberger (1993), Triplett (1996), and Oliner and Sichel (2000) are special cases of the Domar (1961) aggregation scheme.

[34] See Jorgenson, Gollop, and Fraumeni (1987), pp. 63-66, 301-322.

percentage points to productivity growth for 1995-2002, compared to 0.23 percentage points for 1989-1995. This reflects the accelerating decline in relative price changes resulting from shortening the product cycle for semiconductors.

2.2.5. Output Growth.

This section presents the sources of GDP growth for the entire period 1948 to 2002. Capital services contribute 1.75 percentage points, labor services 1.05 percentage points, and productivity growth only 0.67 percentage points. Input growth is the source of nearly 80.6 percent of U.S. growth over the past half century, while productivity has accounted for 19.4 percent. Figure 2.11 shows the relatively modest contributions of productivity in all sub-periods.

Almost four-fifths of the contribution of capital reflects the accumulation of capital stock, while improvement in the quality of capital accounts for about one-fifth. Similarly, increased labor hours account for 68 percent of labor's contribution; the remainder is due to improvements in labor quality. Substitutions among capital and labor inputs in response to price changes are essential components of the sources of economic growth.

A look at the U.S. economy before and after 1973 reveals familiar features of the historical record. After strong output and productivity growth in the 1950's, 1960's and early 1970's, the U.S. economy slowed markedly through 1989, with output growth falling from 3.78 percent to 3.06 percent and productivity growth declining from 0.80 percent to 0.38 percent. The contribution of capital input also slowed from 1.94 percent for 1948-73 to 1.53 percent for 1973-89. This contributed to sluggish ALP growth -- 2.72 percent for 1948-73 and 1.46 percent for 1973-89.

Relative to the period 1989-1995, output growth increased by 1.50 percent in 1995-2002. The contribution of IT production jumped by 0.38 percent, relative to 1989-1995, but still accounted for only 21.9 percent of the increased growth of output. Although the contribution of IT has increased steadily throughout the

period 1948-2002, there has been a sharp response to the acceleration in the IT price decline in 1995. Nonetheless, almost three-quarters of the increased output growth can be attributed to non-IT products.

Between 1989-1995 and 1995-2002 the contribution of capital input jumped by 0.80 percentage points, the contribution of labor input declined by 0.10 percent, and productivity accelerated by 0.79 percent. Growth in ALP rose 1.36 percent as more rapid capital deepening and growth in productivity offset slower improvement in labor quality. Growth in hours worked slowed as labor markets tightened considerably, even as labor force participation rates increased.[35]

The contribution of capital input reflects the investment boom of the late 1990's as businesses, households, and governments poured resources into plant and equipment, especially computers, software, and communications equipment. The contribution of capital, predominantly IT, is considerably more important than the contribution of labor. The contribution of IT capital services has grown steadily throughout the period 1948-2002, but Figure 2.6 reflects the impact of the accelerating decline in IT prices.

After maintaining an average rate of 0.38 percent for the period 1973-89, productivity growth declined to 0.35 percent for 1989-95 and then vaulted to 1.14 percent per year for 1995-2002. This is a major source of growth in output and ALP for the U.S. economy (Figures 2.11 and 2.12). Productivity growth for 1995-2002 is considerably higher than the rate of 1948-73 and the U.S. economy is recuperating from the anemic productivity growth of the past two decades. Although less than half of the acceleration in productivity from 1989-5 to 1995-2002 can be attributed to IT production, this is far greater than the 5.01 percent share of IT in the GDP in 2002.

---

[35] Lawrence Katz and Alan Krueger (1999) analyze the recent performance of the U.S. labor market.

2.2.6. Average Labor Productivity.

Output growth is the sum of growth in hours and average labor productivity. Table 2.7 shows the breakdown between growth in hours and ALP for the same periods as in Table 2.6. For the period 1948-2002, ALP growth predominated in output growth, increasing 2.23 percent per year, while hours worked increased 1.23 percent per year. As shown above, ALP growth depends on capital deepening, a labor quality effect, and overall productivity growth.

Figure 2.12 reveals the well-known productivity slowdown of the 1970's and 1980's, emphasizing the sharp acceleration in labor productivity growth in the late 1990's. The slowdown through 1989 reflects reduced capital deepening, declining labor quality growth, and decelerating growth in total factor productivity. The growth of ALP recovered slightly during the early 1990's with a slump in capital deepening more than offset by a revival in labor quality growth and an up-tick in total factor productivity growth. A slowdown in hours combined with middling ALP growth during 1989-1995 to produce a further slide in the growth of output. In previous cyclical recoveries during the postwar period, output growth accelerated during the recovery, powered by more rapid growth of hours and ALP.

Accelerating output growth during 1995-2002 reflects modest growth in labor hours and a sharp increase in ALP growth[36]. Comparing 1989-1995 to 1995-2002, the rate of output growth jumped by 1.50 percent -- due to an increase in hours worked of 0.14 percent and an upward bound in ALP growth of 1.36 percent. Figure 2.12 shows the acceleration in ALP growth is due to capital deepening as well as faster total factor productivity growth. Capital deepening contributed 0.74 percentage points, counterbalancing a negative contribution of labor quality of 0.16 percent. The acceleration in total factor productivity growth added 0.79 percentage points.

[36] Stiroh (2002) shows that ALP growth is concentrated in IT-producing and IT-using industries.

2.2.7. Research Opportunities.

The use of computers, software, and communications equipment must be carefully distinguished from the production of IT[37]. Massive increases in computing power, like those experienced by the U.S. economy, have two effects on growth. First, as IT producers become more efficient, more IT equipment and software is produced from the same inputs. This raises productivity in IT-producing industries and contributes to productivity growth for the economy as a whole. Labor productivity also grows at both industry and aggregate levels.

Second, investment in information technology leads to growth of productive capacity in IT-using industries. Since labor is working with more and better equipment, this increases ALP through capital deepening. If the contributions to aggregate output are captured by capital deepening, aggregate productivity growth is unaffected[38]. Increasing deployment of IT affects productivity growth only if there are spillovers from IT-producing industries to IT-using industries.

Jorgenson, Ho, and Stiroh (2004) trace the increase in aggregate productivity growth to its sources in individual industries. Jorgenson and Stiroh (2000a, 2000b) present the appropriate methodology and preliminary results. Stiroh (2000) shows that aggregate ALP growth can be attributed to productivity growth in IT-producing and IT-using industries.

2.3. Demise of Traditional Growth Accounting.

2.3.1. Introduction.

The early 1970's marked the emergence of a rare professional consensus on economic growth, articulated in two strikingly dissimilar books. Kuznets

---

[37] Economics and Statistics Administration (2000), Table 3.1, p. 23, lists IT-producing industries.

[38] Baily and Gordon (1988).

summarized his decades of empirical research in *Economic Growth of Nations* (1971). "[39] Solow's book *Economic Growth* (1970), modestly subtitled "An Exposition", contained his 1969 Radcliffe Lectures at the University of Warwick. In these lectures Solow also summarized decades of theoretical research, initiated by the work of Roy Harrod (1939) and Domar (1946).[40]

Let me first consider the indubitable strengths of the perspective on growth that emerged victorious over its many competitors in the early 1970's. Solow's neo-classical theory of economic growth, especially his analysis of steady states with constant rates of growth, provided conceptual clarity and sophistication. Kuznets generated persuasive empirical support by quantifying the long sweep of historical experience of the United States and thirteen other developed economies. He combined this with quantitative comparisons among a developed and developing economies during the postwar period.

With the benefit of hindsight the most obvious deficiency of the traditional framework of Kuznets and Solow was the lack of a clear connection between the theoretical and the empirical components. This lacuna can be seen most starkly in the total absence of cross references between the key works of these two great economists. Yet they were working on the same topic, within the same framework, at virtually the same time, and in the very same geographical Location --*Cambridge, Massachusetts*!

Searching for analogies to describe this remarkable coincidence of views on growth, we can think of two celestial bodies on different orbits,

---

[39]The enormous impact of this research was recognized in the same year by the Royal Swedish Academy of Sciences in awarding the third Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel to Kuznets "for his empirically founded interpretation of economic growth which has led to new and deepened insight into the economic and social structure and process of development." See Assar Lindbeck (1992), p. 79.

[40]Solow's seminal role in this research, beginning with his brilliant and pathbreaking essay of 1956, "A Contribution to the Theory of Economic Growth", was recognized, simply and elegantly, by the Royal Swedish Academy of Sciences inawarding Solow the Nobel Prize in Economics in 1987 "for his contributions to the theory of economic growth." See Karl-Goran Maler (1992), p. 191. Solow (1999) presents an updated version of his exposition of growth theory.

momentarily coinciding from our earth-bound perspective at a single point in the sky and glowing with dazzling but transitory luminosity. The indelible image of this extraordinary event has been burned into the collective memory of economists, even if the details have long been forgotten. The resulting professional consensus, now obsolete, remained the guiding star for subsequent conceptual development and empirical observation for decades.

2.3.2. Human Capital.

The initial challenge to the framework of Kuznets and Solow was posed by Denison's magisterial study, *Why Growth Rates Differ* (1967). Denison retained NNP as a measure of national product and capital stock as a measure of capital input, adhering to the conventions employed by Kuznets and Solow. Denison's comparisons among nine industrialized economies over the period 1950-1962 were cited extensively by both Kuznets and Solow.

However, Denison departed from the identification of labor input with hours worked by Kuznets and Solow. He followed his earlier study of U.S. economic growth, *The Sources of Economic Growth in the United States and the Alternatives Before Us*, published in 1962. In this study he had constructed constant quality measures of labor input, taking into account differences in the quality of hours worked due to the age, sex, and educational attainment of workers.

Kuznets (1971), recognizing the challenge implicit in Denison's approach to measuring labor input, presented his own version of Denison's findings.[41] He carefully purged Denison's measure of labor input of the effects of changes in educational attainment. Solow, for his part, made extensive references to Denison's findings on the growth of output and capital stock, but avoided a detailed reference to Denison's measure of labor input. Solow adhered instead to

[41]Kuznets (1971), Table 9, part B, pp. 74-75.

hours worked (or "man-hours" in the terminology of the early 1970's) as a measure of labor input.[42]

Kuznets showed that "... with one or two exceptions, the contribution of the factor inputs per capita was a minor fraction of the growth rate of per capita product."[43] For the United States during the period 1929 to 1957, the growth rate of productivity or output per unit of input exceeded the growth rate of output per capita. According to Kuznets' estimates, the contribution of increases in capital input per capita over this extensive period was negative!

2.3.3. Solow's Surprise.

The starting point for our discussion of the demise of traditional growth accounting is a notable but neglected article by the great Dutch economist Jan Tinbergen (1942), published in German during World War II. Tinbergen analyzed the sources of U.S. economic growth over the period 1870-1914. He found that efficiency accounted only a little more than a quarter of growth in output, while growth in capital and labor inputs accounted for the remainder. This was precisely the opposite of the conclusion that Kuznets (1971) and Solow (1970) reached almost three decades later!

The notion of efficiency or "total factor productivity" was introduced independently by George Stigler (1947) and became the starting point for a major research program at the National Bureau of Economic Research. This program employed data on output of the U.S. economy from earlier studies by the National Bureau, especially the pioneering estimates of the national product by Kuznets (1961). The input side employed data on capital from Raymond Goldsmith's (1962) system of national wealth accounts. However, much of the data was generated by John Kendrick (1956, 1961), who employed an explicit

---

[42]Solow (1970), pp. 2-7. However, Solow (1988), pp. 313-314, adopted Denison's perspective on labor input in his Nobel Prize address. At about the same time this view was endorsed by Becker (1993a), p. 24, in his 1989 Ryerson Lecture ad the University of Chicago. Becker (1993b) also cited Denison in his Nobel Prize address.

[43]Kuznets (1971), p. 73.

system of national production accounts, including measures of output, input, and productivity for national aggregates and individual industries.[44]

The econometric models of Paul Douglas (1948) and Tinbergen were integrated with data from the aggregate production accounts generated by Abramovitz (1956) and Kendrick (1956) in Solow's justly celebrated 1957 article, "Technical Change and the Aggregate Production Function". Solow identified "technical change" with shifts in the production function. Like Abramovitz, Kendrick, and Kuznets, he attributed almost all of U.S. economic growth to "residual" growth in productivity.[45]

Kuznets' (1971) international comparisons strongly reinforced the findings of Abramovitz (1956), Kendrick (1956), and Solow (1957), which were limited to the United States.[46] According to Kuznets, economic growth was largely attributable to the Solow residual between the growth of output and the growth of capital and labor inputs, although he did not use this terminology. Kuznets' assessment of the significance of his empirical conclusions was unequivocal:

> (G)iven the assumptions of the accepted national economic accounting framework, and the basic demographic and institutional processes that control labor supply, capital accumulation, and initial capital-output ratios, this major conclusion -- that the distinctive feature of modern economic growth, the high rate of growth of per capita product is for the most part attributable to a high rate of growth in productivity -- is inevitable.[47]

---

[44]Updated estimates based on Kendrick's framework are presented by Kendrick (1973) and Kendrick and Grossman (1980).

[45]This finding is called "Solow's Surprise" by William Easterly (2001) and is listed as one of the "stylized facts" about economic growth by Robert King and Sergio Rebelo (1999).

[46]A survey of international comparisons, including Tinbergen (1942)and Kuznets (1971), is given in my paper with Christensen and Cummings (1980), presented at the forty-fourth meeting of the Conference on Research and Wealth, held at Williamsburg, Virginia, in 1975.

[47]Kuznets (1971), p. 73; see also, pp. 306-309.

The empirical findings summarized by Kuznets have been repeatedly corroborated in investigations that employ the traditional approach to growth accounting. This approach identifies output with real NNP, labor input with hours worked, and capital input with real capital stock.[48] Kuznets (1978) interpreted the Solow residual as due to exogenous technological innovation. This is consistent with Solow's (1957) identification of the residual with technical change. Successful attempts to provide a more convincing explanation of the Solow residual have led, ultimately, to the demise of the traditional framework.[49]

2.3.4. Radical Departure.

The most serious challenge to the traditional approach growth accounting was presented in my 1967 paper with Griliches, "The Explanation of Productivity Change". Griliches and I departed far more radically than Denison from the measurement conventions of Kuznets and Solow. We replaced NNP with GNP as a measure of output and introduced constant quality indexes for both capital and labor inputs.

The key idea underlying our constant quality index of labor input, like Denison's, was to distinguish among different types of labor inputs. We combined hours worked for each type into a constant quality index of labor input, using the index number methodology Griliches (1960) had developed for U.S. agriculture. This considerably broadened the concept of substitution employed by Solow (1957). While he had modeled substitution between capital and labor inputs, Denison, Griliches and I extended the concept of substitution to include

---

[48]For recent examples, see Michael Dertouzos, Solow, and Richard Lester (1989) and Hall (1988, 1990a).

[49]A detailed survey of research on sources of economic growth is given in my 1990 article, "Productivity and Economic Growth", presented at the The Jubilee of the Conference on Research in Income and Wealth, held in Washington, D.C., in 1988, commemorating the fiftieth anniversary of the founding of the Conference by Kuznets. More recent surveys are presented in Griliches' (2000) posthumous book, *R&D, Education, and Productivity,* and Charles Hulten's (2001) article, "Total Factor Productivity: A Short Biography".

different types of labor inputs as well. This altered, irrevocably, the allocation of economic growth between substitution and technical change.[50]

Griliches and I introduced a constant quality index of capital input by distinguishing among types of capital inputs. To combine different types of capital into a constant quality index, we identified the prices of these inputs with rental prices, rather than the asset prices used in measuring capital stock. For this purpose we used a model of capital as a factor of production I had introduced in my 1963 article, "Capital Theory and Investment Behavior". This made it possible to incorporate differences among depreciation rates on different assets, as well as variations in returns due to the tax treatment of different types of capital income, into our constant quality index of capital input.[51]

Finally, Griliches and I replaced the aggregate production function employed by Denison, Kuznets, and Solow with the production possibility frontier introduced in my 1966 paper, "The Embodiment Hypothesis". This allowed for joint production of consumption and investment goods from capital and labor inputs. I had used this approach to generalize Solow's (1960) concept of embodied technical change, showing that economic growth could be interpreted, equivalently, as "embodied" in investment or "disembodied" in productivity growth. My 1967 paper with Griliches removed this indeterminacy by introducing constant quality price indexes for investment goods.[52]

---

[50]Constant quality indexes of labor input are discussed detail by Jorgenson, Gollop, and Fraumeni (1987), Chapters 3 and 8, pp. 69-108 and 261-300, and Jorgenson, Ho, and Stiroh (2004).

[51]I have presented a detailed survey of empirical research on the measurement of capital input in my 1989 paper, "Capital as a Factor of Production". Earlier surveys were given in my 1973 and 1980 papers and Diewert's (1980) contribution to the forty-fifth meeting of the Conference on Income and Wealth, held at Toronto, Ontario, in 1976. Hulten (1990) surveyed conceptual aspects of capital measurement in his contribution to the Jubilee of the Conference on Research in Income and Wealth in 1988.

[52]As a natural extension of Solow's (1956) one-sector neo-classical model of economic growth, his 1960 model of embodiment had only a single output and did

Griliches and I showed that changes in the quality of capital and labor inputs and the quality of investment goods explained most of the Solow residual. We estimated that capital and labor inputs accounted for eighty-five percent of growth during the period 1945-1965, while only fifteen percent could be attributed to productivity growth. Changes in labor quality explained thirteen percent of growth, while changes in capital quality another eleven percent.[53] Improvements in the quality of investment goods enhanced the growth of both investment goods output and capital input; the net contribution was only two percent of growth.[54]

2.3.5. The Rees Report.

The demise of the traditional framework for productivity measurement began with the Panel to Review Productivity Statistics of the National Research Council, chaired by Albert Rees. The Rees Report of 1979, *Measurement and Interpretation of Productivity*, became the cornerstone of a new measurement framework for the official productivity statistics. This was implemented by the Bureau of Labor Statistics (BLS), the U.S. government agency responsible for these statistics.

Under the leadership of Jerome Mark and Edwin Dean the BLS Office of Productivity and Technology undertook the construction of a production account for the U.S. economy with measures of capital and labor inputs and total factor

---

not allow for the introduction of a separate price index for investment goods. Recent research on Solow's model of embodiment is surveyed by Greenwood and Boyan Jovanovic (2001) and discussed by Solow (2001). Solow's model of embodiment is also employed by Whelan (2002).

[53]See Jorgenson and Griliches (1967), Table IX, p. 272. We also attributed thirteen percent of growth to the relative utilization of capital, measured by energy consumption as a proportion of capacity; however, this is inappropriate at the aggregate level, as Denison (1974), p. 56, pointed out. For additional details, see Jorgenson, Gollop, and Fraumeni (1987), especially pp. 179-181.

[54]Using Gordon's (1990) estimates of improvements in the quality of producers' durables, Hulten (1992b) estimated this proportion as 8.5 percent of the growth of U.S. manufacturing output for the period 1949-1983.

productivity, renamed multifactor productivity.[55] The BLS (1983) framework was based on GNP rather than NNP and included a constant quality index of capital input, displacing two of the key conventions of the traditional framework of Kuznets and Solow.[56]

However, BLS retained hours worked as a measure of labor input until July 11, 1994, when it released a new multifactor productivity measure including a constant quality index of labor input as well. Meanwhile, BEA (1986) had incorporated a constant quality price index for computers into the national accounts -- over the strenuous objections of Denison (1989). This index was incorporated into the BLS measure of output, completing the displacement of the traditional framework of economic measurement by the conventions employed in my papers with Griliches.[57]

The official BLS (1994) estimates of multifactor productivity have over-turned the findings of Abramovitz (1956) and Kendrick (1956), as well as those of Kuznets (1971) and Solow (1970). The official statistics have corroborated the findings summarized in my 1990 survey paper, "Productivity and Economic Growth". These statistics are now consistent with the original findings of Tinbergen (1942), as well as my paper with Griliches (1967), and the results I have presented in Section 2.2.

The approach to growth accounting presented in my 1987 book with Gollop and Fraumeni and the official statistics on multifactor productivity published by the BLS in 1994 has now been recognized as the international standard. The new framework for productivity measurement is presented in *Measuring Productivity*, a Manual published by the Organisation for Economic Co-Operation and Development (OECD) and written by Schreyer (2001). The expert advisory group

---

[55]A detailed history of the BLS productivity measurement program is presented by Dean and Harper (2001).

[56]The constant quality index of capital input became the international standard for measuring productivity in Blades' (2001) OECD manual, *Measuring Capital*.

[57]The constant quality index of labor input became the international standard in the United Nations (1993) *System of National Accounts*.

for this manual was chaired by Dean, former Associate Commissioner for Productivity at the BLS, and leader of the successful effort to implement the Rees Report (1979).

3. International Comparisons

3.1. Introduction.

In this section I present international comparisons of economic growth among the G7 nations – Canada, France, Germany, Italy, Japan, the U.K., and the U.S. These comparisons focus on the impact of investment in information technology (IT) equipment and software over the period 1980-2001. In 1998 the G7 nations accounted for nearly sixty percent of world output[58] and a much larger proportion of world investment in IT. Economic growth in the G7 has experienced a strong revival since 1995, driven by a powerful surge in IT investment.

The resurgence of economic growth in the United States during the 1990's and the crucial role of IT investment has been thoroughly documented and widely discussed.[59] Similar trends in the other G7 economies have been more difficult to detect, partly because of discrepancies among official price indexes for IT equipment and software identified by Andrew Wyckoff.[60] Paul Schreyer has constructed "internationally harmonized" IT prices that eliminate many of these discrepancies.[61]

Using internationally harmonized prices, I have analyzed the role of investment and productivity as sources of growth in the G7 countries over the period 1980-2001. I have subdivided the period in 1989 and 1995 in order to

---

[58]See Angus Maddison (2001) for 1998 data for world GDP and the GDP of each of the G7 countries.
[59]See Jorgenson and Stiroh (2000) and Oliner and Sichel (2000).
[60]See Wyckoff (1995)
[61]See Schreyer (2000). Alessandra Colecchia and Schreyer (2002) have employed these internationally harmonized prices in measuring the impact of IT investment.

focus on the most recent experience. I have decomposed growth of output for each country between growth of input and productivity. Finally, I have allocated the growth of input between investments in tangible assets, especially information technology and software, and human capital.

Growth in IT capital input per capita jumped to double-digit levels in the G7 nations after 1995. This can be traced to acceleration in the rate of decline of IT prices, analyzed in my Presidential Address to the American Economic Association.[62] The powerful surge in investment was most pronounced in Canada, but capital input growth in Japan, the U.S., and the U.K. was only slightly lower. France, Germany, and Italy also experienced double-digit growth, but lagged considerably behind the leaders.

During the 1980's productivity played a minor role as a source of growth for the G7 countries except Japan, where productivity accounted for thirty percent of economic growth. Productivity accounted for only sixteen percent of growth in the U.S., thirteen percent in France, twelve percent in the U.K., and eleven percent in Germany; only two percent of growth in Canada was due to productivity, while the decline of productivity retarded growth by fourteen percent in Italy. Between 1989 and 1995 productivity growth declined further in the G7 nations, except for Italy and Germany. Productivity declined for France and the U.K. but remained positive for the U.S., Canada, and Japan.

Productivity growth revived in all the G7 countries after 1995, again with the exception of Germany and Italy. The resurgence was most dramatic in Canada, The U.K., and France, partly offsetting years of dismal productivity growth. Japan exhibited the highest growth in output per capita among the G7 nations from 1980 to 1995. Japan's level of output per capita rose from the lowest in the G7 to the middle of the group. Although this advance owed more to input per capita than productivity, Japan's productivity growth far outstripped the other

[62]See Jorgenson (2001).

members of the G7. Nonetheless, Japan's productivity remained the lowest among the G7 nations.

The U.S. led the G7 in output per capita for the period 1989-2000. Canada's edge in output per capita in 1980 had disappeared by 1989. The U.S. led the G7 countries in input per capita during 1980-2000, but U.S. productivity languished below the levels of Canada, France, and Italy.

In Section 3.2 I outline the methodology for this study, based on Section 2 above. I have revised and updated the U.S. data presented there through 2001. Comparable data on investment in information technology have been have been constructed for Canada by Statistics Canada.[63] Data on IT for France, Germany, Italy, and the U.K. have been developed for the European Commission by Bart Van Ark, *et al*.[64] Finally, data for Japan have been assembled by myself and Kazuyuki Motohashi for the Research Institute on Economy, Trade, and Industry.[65] I have linked these data by means of the OECD's purchasing power parities for 1999.[66]

In Section 3.3 I consider the impact of IT investment and the relative importance of investment and productivity in accounting for economic growth among the G7 nations. Investments in human capital and tangible assets, especially IT equipment and software, account for the overwhelming proportion of growth. Differences in the composition of capital and labor inputs are essential for identifying persistent international differences in output and accounting for the impact of IT investment.

In Section 3.4 I consider alternative approaches to international comparisons. The great revival of interest in economic growth among economists dates from Maddison's (1982) updating and extension of Simon Kuznets' (1971) long-term estimates of the growth of national product and population for

---

[63]See John Baldwin and Tarek Harchaoui (2002).
[64]See Van Ark, Johanna Melka, Nanno Mulder, Marcel Timmer, and Gerard Ypma (2002).
[65]See Jorgenson and Motohashi (2003)
[66]See OECD (2002). Current data on purchasing power parities are available from the OECD website: http://www.sourceoecd.org.

fourteen industrialized countries, including the G7 nations. Maddison (1982, 1991) added Austria and Finland to Kuznets' list and presented growth rates covering periods beginning as early as 1820 and extending through 1989.

Maddison (1987, 1991) also generated growth accounts for major industrialized countries, but did not make level comparisons like those presented in Section 3.2 below. As a consequence, productivity differences were omitted from the canonical formulation of "growth regressions" by William Baumol (1986). This proved to be a fatal flaw in Baumol's regression model, remedied by Nazrul Islam's (1995) panel data model. Section 3.5 concludes.

3.2. Investment and Productivity.

My papers with Laurits Christensen and Dianne Cummings (1980, 1981) developed growth accounts for the United States and its major trading partners -- Canada, France, Germany, Italy, Japan, Korea, The Netherlands, and the United Kingdom for 1947-1973. We employed GNP as a measure of output and incorporated constant quality indices of capital and labor input for each country. Our 1981 paper compared levels of output, inputs, and productivity for all nine nations.

I have updated the estimates for the G7 - Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States - through 1995 in earlier work. The updated estimates are presented in my papers with Chrys Dougherty (1996, 1997) and Eric Yip (2000). We have shown that productivity accounted for only eleven percent of economic growth in Canada and the United States over the period 1960-1995.

My paper with Yip (2000) attributed forty-seven percent of Japanese economic growth during the period 1960-1995 to productivity growth. The proportion attributable to productivity approximated forty percent of growth for the four European countries – France (.38), Germany (.42), Italy (.43), and the United Kingdom (.36). Input growth predominated over productivity growth for all the G7 nations.

I have now incorporated new data on investment in information technology equipment and software for the G7. I have also employed internationally harmonized prices like those constructed by Schreyer (2000). As a consequence, I have been able to separate the contribution of capital input to economic growth into IT and Non-IT components. While IT investment follows similar patterns in all the G7 nations, Non-IT investment varies considerably and helps to explain important differences in growth rates among the G7.

3.2.1. Comparisons of Output, Input, and Productivity.

My first objective is to extend my estimates for the G7 nations with Christensen, Cummings, Dougherty, and Yip to the year 2001. Following the methodology of my Presidential Address, I have chosen GDP as a measure of output. I have included imputations for the services of consumers' durables as well as land, buildings, and equipment owned by nonprofit institutions. I have also distinguished between investments in information technology equipment and software and investments in other forms of tangible assets.

A constant quality index of capital input is based on weights that reflect differences in capital consumption, tax treatment, and the rate of decline of asset prices. I have derived estimates of capital input and property income from national accounting data. Similarly, a constant quality index of labor input is based on weights by age, sex, educational attainment, and employment status. I have constructed estimates of hours worked and labor compensation from labor force surveys for each country.

In Table 3.1 I present output per capita for the G7 nations from 1980 to 2001, taking the U.S. as 100.0 in 2000. Output and population are given separately in Tables 3.2 and 3.3. I use 1999 purchasing power parities from the OECD to convert output from domestic prices for each country into U.S. dollars. The U.S. gained the lead among the G7 countries in output per capita after 1995. Canada led the U.S. in 1980, but fell behind during the 1995. The U.S.-Canada gap widened considerably during the 1990’s.

The four major European nations – the U.K., France, Germany, and Italy - had similar levels of output per capita throughout the period 1980-2001. Japan rose from last place in 1980 to fourth among the G7 in 2001, lagging considerably behind the U.S. and Canada, but only slightly behind the U.K. in 2001. Japan led the G7 in the growth of output per capita from 1980-1995, but fell behind the U.S., Canada, the U.K., France, and Italy after 1995.

In Table 3.1 I present input per capita for the G7 over the period 1980-2000, taking the U.S. as 100.0 in 2000. I express input per capita in U.S. dollars, using purchasing power parities constructed for this study.[67] The U.S. was the leader among the G7 in input per capita throughout the period. In 2001 Canada ranked next to the U.S. with Japan third and Germany fourth. France and Italy started at the bottom of the ranking and remained there throughout the period.

In Table 3.1 I also present productivity levels for the G7 over the period 1980-2001. Productivity is defined as the ratio of output to input, including both capital and labor inputs. Italy led in 1980 and Canada was the productivity leader throughout the period 1989-2001 with France close behind. Japan made substantial gains in productivity during the period, while there were more modest increases in the U.S., Canada, the U.K., France, and Germany, and a decline in Italy.

I summarize growth in output and input per capita and productivity for the G7 nations in Table 3.4. I present growth rates of output and population for the period 1980-2001 in Tables 3.2 and 3.3. Output growth slowed in the G7 after 1989, but revived for all nations except Japan and Germany after 1995. Output per capita followed a similar pattern with Canada barely expanding during the period 1990-1995.

---

[67]The purchasing power parities for outputs are based on OECD (2002). Purchasing power parities for inputs follow the methodology described in detail by Jorgenson and Yip (2001).

Japan led in growth of output and output per capita through 1995, but fell to the lower echelon of the G7 after 1995. Japan also led in productivity growth throughout the period 1980-2001. For all countries and all time periods, except for Germany during the period 1989-1995 and Japan after 1995, the growth of input per capita exceeded growth of productivity by a substantial margin. Productivity growth in the G7 slowed during the period 1989-1995, except for Germany and Italy, where productivity slumped after 1995.

Italy led the G7 in growth of input per capita for the periods 1980-1989 and 1995-2001, but relinquished leadership to the U.K. for the period 1989-1995. Differences among input growth rates are smaller than differences among output growth rates, but there was a slowdown in input growth during 1989-1995 throughout the G7. After 1995 growth of input per capita increased in every G7 nation except Japan.

3.2.2. Comparisons of Capital and Labor Quality.

A constant quality index of capital input weights capital inputs by property compensation per unit of capital. By contrast an index of capital stock weights different types of capital by asset prices. The ratio of capital input to capital stock measures the average quality of a unit of capital. This represents the difference between the constant quality index of capital input and the index of capital stock employed, for example, by Kuznets (1971) and Robert Solow (1970).

In Table 3.5 I present capital input per capita for the G7 countries over the period 1980-2001 relative to the U.S. in 2000. The U.S. was the leader in capital input per capita throughout the period, while the U.K. was the laggard. Canada led the remaining six countries in 1980, but was overtaken by Germany and Italy in 1995. Italy led the rest of the G7 through 2001, but lagged considerably behind the United States.

The picture for capital stock per capita has some similarities to capital input, but there are important differences. Capital stock levels do not

accurately reflect the substitutions among capital inputs that accompany investments in tangible assets, especially investments in IT equipment and software. The U.S. led the G7 in capital stock per capita as well as capital input after 1989, while Japan led in 1980 and was second to the U.S. after 1989. The U.K. lagged the remaining countries of the G7 throughout the period.

The behavior of capital quality highlights the differences between the constant quality index of capital input and capital stock. There are important changes in capital quality over time and persistent differences among countries, so that heterogeneity in capital input must be taken into account in international comparisons of economic performance. Canada was the international leader in capital quality throughout the period 1980-2001, while Japan ranked at the bottom of the G7.

I summarize growth in capital input and capital stock per capita, as well as capital quality for the G7 nations in Table 3.8. Italy was the international leader in capital input growth from 1980-1989, while Canada was the laggard. The U.K. led from 1989-1995, while Canada lagged considerably behind the rest of the G7. The U.S. took the lead after 1995. There was a slowdown in capital input growth throughout the G7 after 1989, except for the U.K., and a revival after 1995 in the U.S., Canada, France, and Italy.

A constant quality index of labor input weights hours worked for different categories by labor compensation per hour. An index of hours worked fails to take quality differences into account. The ratio of labor input to hours worked measures the average quality of an hour of labor, as reflected in its marginal product. This represents the difference between the constant quality index of labor input and the index of hours worked employed, for example, by Kuznets (1971) and Solow (1970).

In Table 3.11 I present labor input per capita for the G7 nations for the period 1980-2001 relative to the U.S. in 2000. Japan was the international leader throughout the period and France and Italy the laggards. Labor input in

Japan was nearly double that of Italy. The U.S. led the remaining G7 nations throughout the period. The U.K. ranked third among the G7 through 1995. Italy and France lagged behind the rest of the G7 for the entire period.

The picture for hours worked per capita has some similarities to labor input, but there are important differences. Japan was the international leader in hours worked per capita. The U.S., Canada, and the U.K. moved roughly in parallel. The U.K. ranked second in 1980 and 1989, while the U.S. ranked second in 1995 and 2001. France and Italy lagged the rest of the G7 from 1980-2001.

The behavior of labor quality highlights the differences between labor input and hours worked. Germany was the leader in labor quality throughout the period 1980-2001 with the U.S. close behind. Canada, the U.K., France, and Japan had similar levels of labor quality throughout the period, but fell short of German and U.S. levels. Italy was the laggard among the G7 in labor quality.

I summarize growth in labor input and hours worked per capita, as well as labor quality for the period 1980-2001 in Table 3.12. Canada and Japan led the G7 nations in labor input growth during the 1980's, France led from 1989-1995 but relinquished its leadership to Italy after 1995. Labor input growth was negative for France during the 1980's, for the U.K., Germany, Italy, and Japan during the period 1989-1995, and for Japan after 1995.

Hours worked per capita fell continuously throughout the period 1980-2001 for Japan and declined for all the G7 nations during the period 1989-1995. Growth in labor quality was positive for the G7 nations in all time periods. Japan was the leader during the 1980's, relinquishing its lead to France during the early 1990's and Italy in the late 1990's. Growth in labor quality and hours worked are equally important as sources of growth in labor input for the G7.

3.3. Investment in Information Technology.

Using data from Tables 3.1 and 3.2, I can assess the relative importance of investment and productivity as sources of economic growth for the G7 nations. Investments in tangible assets and human capital greatly predominated over

productivity during the period 1980-2001. While productivity fell in Italy during this period, the remaining G7 countries had positive productivity growth for the period as a whole.

Similarly, using data from Table 3.5 I can assess the relative importance of growth in capital stock and capital quality. Capital input growth was positive for all countries for the period 1980-2001 and all three sub-periods. Capital quality growth was positive for the period as a whole for all G7 countries. Although capital stock predominated in capital input growth, capital quality was also quantitatively significant, especially after 1995.

Finally, using data from Table 3.11 I can assess the relative importance of growth in hours worked and labor quality. Hours worked per capita declined for France, Germany, and Japan, while labor quality rose in these nations during the period 1980-2001. For the U.S., Canada, the U.K., and Italy, both hours worked per capita and labor quality rose. I conclude that labor quality growth is essential to the analysis of growth in labor input.

3.3.1. Investment in IT Equipment and Software

The final step in the comparison of patterns of economic growth among the G7 nations is to analyze the impact of investment in information technology equipment and software. In Table 3.6 I present levels of IT capital input per capita for the G7 for the period 1980-2001, relative to the U.S. in 2000. The U.S. overtook Germany in 1989 and remained the leader through 2001. Canada and Japan lagged behind the rest of the G7 through 1995, but France fell into last place in 2001.

Table 3.6 reveals substantial differences between IT capital stock and IT capital input. The G7 nations began with very modest stocks of IT equipment and software per capita in 1980. These stocks expanded rapidly during the period 1980-2001. The U.S. led in IT capital stock throughout the period, while Japan moved from the third lowest level in 1980 to the second highest in 2001.

IT capital quality reflects differences in the composition of IT capital input, relative to IT capital stock. A rising level of capital quality indicates a shift toward short-lived assets, such as computers and software. This shift is particularly dramatic for the U.S., Canada, and Japan, while the composition of IT capital stock changed relatively less for the U.K., France, Germany, and Italy. Patterns for Non-IT capital input, capital stock, and capital quality largely reflect those for capital as a whole, presented in Table 3.5.

I give growth rates for IT capital input per capita, capital stock per capita, and capital quality in Table 3.9. The G7 nations have exhibited double-digit growth in IT capital input per capita since 1995. Canada was the international leader during this period with Japan close behind. Japan was the leader in growth of IT capital input during the 1980's, another period of double-digit growth in the G7. However, Japanese IT growth slowed substantially during 1989-1995 and Canada gained the lead.

Patterns of growth for IT capital stock per capita are similar to those for IT capital input for the four European countries. Changes in the composition of IT capital stock per capita were important sources of growth of IT capital input per capita for the U.S., Canada, and Japan. IT capital stock also followed the pattern of IT capital input with substantial growth during the 1980's, followed by a pronounced lull during the period 1989-1995. After 1995 the growth rates of IT capital stock surged in all the G7 countries, except Germany, but exceeded the rates of the 1980's only for the U.S. and Japan.

Finally, growth rates for IT capital quality reflect the rates at which shorter-lived IT assets are substituted for longer-lived assets.
Japan led in the growth of capital quality during the 1980's, but relinquished its lead to Canada in 1989. IT capital quality growth for the Canada substantially outstripped that of the remaining G7 countries for the period 1989-2001. Patterns of growth in Non-IT capital input per capita, Non-IT capital

stock per capita, and Non-IT capital quality given in Table 3.10 largely reflect those for capital as a whole presented in Table 3.8.

Table 3.13 and Figure 3.1 present the contribution of capital input to economic growth for the G7 nations, divided between IT and Non-IT. The powerful surge of IT investment in the U.S. after 1995 is mirrored in similar jumps in growth rates of the contribution of IT capital through the G7. The contribution of IT capital input was similar during the 1980's and the period 1989-1995 for all the G7 nations, despite the dip in rates of economic growth after 1989. Japan is an exception to this general pattern with a contribution of IT capital comparable to that of the U.S. during the 1980's, followed by a decline in this contribution from 1989-1995, reflecting the sharp downturn in Japanese economic growth.

The contribution of Non-IT capital input to economic growth after 1995 exceeded that for IT capital input for four of the G7 nations; the exceptions were Canada, the U.K., and Japan. The U.S. stands out in the magnitude of the contribution of capital input after 1995. Both IT and Non-IT capital input contributed to the U.S. economic resurgence of the last half of the 1990's. Despite the strong performance of IT investment in Japan after 1995, the contribution of capital input declined substantially; this contribution also declined for the U.K. and Germany.

3.3.2. The Relative Importance of Investment and Productivity.

Table 3.14 and Figure 3.2 present contributions to economic growth from productivity, divided between the IT-producing and Non-IT-producing industries. The methodology for this division follows Triplett (1996). The contribution of IT-producing industries is positive throughout the period 1980-2001 and jumps substantially after 1995. Since the level of productivity in Italy is higher in 1980 than in 2001, it is not surprising that the contribution of productivity growth in the Non-IT industries was negative throughout the period. Productivity

in these industries also declined during 1989-1995 in Canada, the U.K., and France and after 1989 in Germany as well as Italy.

Table 3.15 and Figure 3.3 give a comprehensive view of the sources of economic growth for the G7. The contribution of capital input alone exceeds that of productivity for most nations and most time periods. The contribution of Non-IT capital input predominates over IT capital input for most countries and most time periods with Canada in 1989-2001, and the U.K. and Japan after 1995 as exceptions. This can be attributed to the unusual weakness in the growth of aggregate demand in these countries. The contribution of labor input varies considerably among the G7 nations with negative contributions after 1995 in Japan, during the 1980's in France, and during the period 1989-1995 in the U.K. and Germany.

Finally, Table 3.16 and Figure 3.4 translate sources of growth into sources of growth in average labor productivity (ALP). ALP, defined as output per hour worked, must be carefully distinguished from overall productivity, defined as output per unit of both capital and labor inputs. Output growth is the sum of growth in hours worked and growth in ALP. ALP growth depends on the contribution of capital deepening, the contribution of growth in labor quality, and productivity growth.

Capital deepening is the contribution of growth in capital input per hour worked and predominates over productivity as a source of ALP growth for the G7 nations. IT capital deepening predominates over Non-IT capital deepening in the U.S. throughout the period 1980-2001 and in Canada after 1989, the U.K., France, and Japan after 1995. Finally, the contribution of labor quality is positive for all the G7 nations through the period.

3.4. Alternative Approaches

Edward Denison's (1967) pathbreaking volume, *Why Growth Rates Differ*, compared differences in growth rates for national income net of capital consumption per capita for the period 1950-62 with differences of levels in 1960

for eight European countries and the U.S. The European countries were characterized by much more rapid growth and a lower level of national income per capita. However, this association did not hold for all comparisons between the individual countries and the U.S. Nonetheless, Denison concluded:[68]

> Aside from short-term aberrations Europe should be able to report higher growth rates, at least in national income per person employed, for a long time. Americans should expect this and not be disturbed by it.

Maddison (1987, 1991) constructed estimates of aggregate output, input, and productivity growth for France, Germany, Japan, The Netherlands, and the United Kingdom for the period 1870-1987. Maddison (1995) extended estimates for the U.S., the U.K., and Japan backward to 1820 and forward to 1992. He defined output as gross of capital consumption throughout the period and constructed constant quality indices of labor input for the period 1913-1984, but not for 1870-1913.

Maddison employed capital stock as a measure of the input of capital, ignoring the changes in the composition of capital stock that are such an important source of growth for the G7 nations. This omission is especially critical in assessing the impact of investment in information technology. Finally, he reduced the growth rate of the price index for investment by one percent per year for all countries and all time periods to correct for biases like those identified by Wyckoff (1995).

3.4.1. Comparisons without Growth Accounts

Kuznets (1971) provided elaborate comparisons of growth rates for fourteen industrialized countries. Unlike Denison (1967), he did not provide level comparisons. Maddison (1982) filled this lacuna by comparing levels of national product for sixteen countries. These comparisons used

---

[68]See Denison (1967), especially Chapter 21, "The Sources of Growth and the Contrast between Europe and the United States", pp. 296-348.

estimates of purchasing power parities by Irving Kravis, Alan Heston, and Robert Summers (1978).[69]

Maddison (1995) extended his long-term estimates of the growth of national product and population to 56 countries, covering the period 1820-1992. Maddison (2001) updated these estimates to 1998 in his magisterial volume, *The World Economy: A Millennial Perspective*. He provided estimates for 134 countries, as well as seven regions of the world – Western Europe, Western Offshoots (Australia, Canada, New Zealand, and the United States), Eastern Europe, Former USSR, Latin America, Asia, and Africa.

Purchasing power parities have been updated by successive versions of the Penn World Table. A complete list of these tables through Mark 5 is given by Summers and Heston (1991). The current version of the Penn World Table is available on the Center for International Comparisons website at the University of Pennsylvania (CICUP). This covers 168 countries for the period 1950-2000 and represents one of the most significant achievements in economic measurement of the postwar period.[70]

3.4.2. Convergence

Data presented by Kuznets (1971), Maddison, and successive versions of the Penn World Table have made it possible to reconsider the issue of convergence raised by Denison (1967). Moses Abramovitz (1986) was the first to take up the challenge by analyzing convergence of output per capita among Maddison's sixteen countries. He found that convergence characterized the postwar period, while there was no tendency toward convergence before 1914 and during the interwar period. Baumol (1986) formalized these results by running a regression of growth

[69]For details see Maddison (1982), pp. 159-168.
[70]See Heston, Summers, and Aten (2002). The CICUP website is at: http://pwt.econ.upenn.edu/aboutpwt.html.

rate of GDP per capita over the period 1870-1979 on the 1870 level of GDP per capita.[71]

In a highly innovative paper on "Crazy Explanations for the Productivity Slowdown" Paul Romer (1987) derived Baumol's "growth regression" from Solow's (1970) growth model with a Cobb-Douglas production function. Romer's empirical contribution was to extend the growth regressions from Maddison's (1982) sixteen advanced countries to the 115 countries in the Penn World Table (Mark 3). Romer's key finding was an estimate of the elasticity of output with respect to capital close to three-quarters. The share of capital in GNP implied by Solow's model was less than half as great.

Gregory Mankiw, David Romer, and David Weil (1992) defended the traditional framework of Kuznets (1971) and Solow (1970). The empirical part of their study is based on data for 98 countries from the Penn World Table (Mark 4). Like Paul Romer (1987), Mankiw, David Romer, and Weil derived a growth regression from the Solow (1970) model; however, they augmented this by allowing for investment in human capital.

The results of Mankiw, David Romer, and Weil (1992) provided empirical support for the augmented Solow model. There was clear evidence of the convergence predicted by the model; in addition, the estimated elasticity of output with respect to capital was in line with the share of capital in the value of output. The rate of convergence of output per capita was too slow to be consistent with 1970 version of the Solow model, but supported the augmented version.

---

[71]Baumol's "growth regression" has spawned a vast literature, recently summarized by Steven Durlauf and Danny Quah (1999, Ellen McGrattan and James Schmitz (1999), and Islam (2003). Much of this literature is based on data from successive versions of the Penn World Table.

3.4.2. Modeling Productivity Differences.

Finally, Islam (1995) exploited an important feature of the Penn World Table overlooked in prior studies. This panel data set contains benchmark comparisons of levels of the national product at five year intervals, beginning in 1960. This made it possible to test an assumption maintained in growth regressions. These regressions had assumed identical levels of productivity for all countries included in the Penn World Table.

Substantial differences in levels of productivity among countries have been documented by Denison (1967), by my papers with Christensen and Cummings (1981), Dougherty (1996, 1999), and Yip (2000) and in Section 2 above. By introducing econometric methods for panel data Islam (1995) was able to allow for these differences. He corroborated the finding of Mankiw, David Romer, and Weil (1992) that the elasticity of output with respect to capital input coincided with the share of capital in the value of output.

In addition, Islam (1995) found that the rate of convergence of output per capita among countries in the Penn World Table substantiated the *unaugmented* version of the Solow (1970) growth model. In short, "crazy explanations" for the productivity slowdown, like those propounded by Paul Romer (1987, 1994), were unnecessary. Moreover, the model did not require augmentation by endogenous investment in human capital, as proposed by Mankiw, David Romer, and Weil (1992).

Islam concluded that differences in technology among countries must be included in econometric models of growth rates. This requires econometric techniques for panel data, like those originated by Gary Chamberlain (1982), rather than the regression methods of Baumol, Paul Romer, and Mankiw, David Romer, and Weil. Panel data techniques have now superseded regression methods in modeling differences in output per capita.

3.5. Conclusions.

I conclude that a powerful surge in investment in information technology and equipment after 1995 characterizes all of the G7 economies. This accounts for a large portion of the resurgence in U.S. economic growth, but contributes substantially to economic growth in the remaining G7 economies as well. Another significant source of the G7 growth resurgence after 1995 is a jump in productivity growth in IT-producing industries.

For Japan the dramatic upward leap in the impact of IT investment after 1995 was insufficient to overcome downward pressures from deficient growth of aggregate demand. This manifests itself in declining contributions of Non-IT capital and labor inputs. Similar downturns are visible in Non-IT capital input in France, Germany, and especially the U.K. after 1995.

These findings are based on new data and new methodology for analyzing the sources of economic growth. Internationally harmonized prices for information technology equipment and software are essential for capturing differences among the G7 nations. Constant quality indices of capital and labor inputs are necessary to incorporate the impacts of investments in information technology and human capital.

Exploiting the new data and methodology, I have been able to show that investment in tangible assets is the most important source of economic growth in the G7 nations. The contribution of capital input exceeds that of productivity for all countries for all periods. The relative importance of productivity growth is far less than suggested by the traditional methodology of Kuznets (1971) and Solow (1970), which is now obsolete.

The conclusion from Islam's (1995) research is that the Solow (1970) model is appropriate for modeling the endogenous accumulation of tangible assets. It is unnecessary to endogenize human capital accumulation as well. The transition path to balanced growth equilibrium after a change in policies that affects

investment in tangible assets requires decades, while the transition after a change affecting investment in human capital requires as much as a century.

4. Economics on Internet Time.

The steadily rising importance of information technology has created new research opportunities in all areas of economics. Economic historians, led by Chandler (2000) and Moses Abramovitz and Paul David (1999, 2001)[72], have placed the information age in historical context. Abramovitz and David present sources of U.S. economic growth for the nineteenth and twentieth centuries. Their estimates, beginning in 1966, are based on the official productivity statistics published by the Bureau of Labor Statistics (1994).

The Solow (1987) Paradox, that we see computers everywhere but in the productivity statistics[73], has been displaced by the economics of the information age. Computers have now left an indelible imprint on the productivity statistics. The remaining issue is whether the breathtaking speed of technological change in semiconductors differentiates this resurgence from previous periods of rapid growth?

Capital and labor markets have been severely impacted by information technology. Enormous uncertainty surrounds the relationship between equity valuations and future growth prospects of the American economy[74]. One theory attributes rising valuations of equities since the growth acceleration began in 1995 to the accumulation of intangible assets, such as intellectual property and organizational capital. An alternative theory treats the high valuations of technology stocks as a bubble that burst during the year 2000.

---

[72] See also: David (1990, 2000) and Gordon (2000).

[73] Griliches (1994), Brynjolfsson and Shinkyu Yang (1996), and Triplett (1999) discuss the Solow Paradox.

[74] Campbell and Shiller (1998) and Shiller (2000) discuss equity valuations and growth prospects. Michael Kiley (1999), Brynjolfsson and Hitt (2000), and Robert Hall (2000, 2001), present models of investment with internal costs of adjustment.

The behavior of labor markets also poses important puzzles. Widening wage differentials between workers with more and less education has been attributed to computerization of the workplace. A possible explanation could be that high-skilled workers are complementary to IT, while low-skilled workers are substitutable. An alternative explanation is that technical change associated with IT is skill-biased and increases the wages of high-skilled workers relative to low-skilled workers[75].

Finally, information technology is altering product markets and business organizations, as attested by the large and growing business literature[76], but a fully satisfactory model of the semiconductor industry remains to be developed[77]. Such a model would derive the demand for semiconductors from investment in information technology in response to rapidly falling IT prices. An important objective is to determine the product cycle for successive generations of new semiconductors endogenously.

The semiconductor industry and the information technology industries are global in their scope with an elaborate international division of labor[78]. This poses important questions about the American growth resurgence. Where is the evidence of a new economy in other leading industrialized countries? I have shown in Section 3 that the most important explanation is the relative paucity of constant quality price indexes for semiconductors and information technology in national accounting systems outside the U.S.

---

[75] Daron Acemoglu (2002) and Katz (2000) survey the literature on labor markets and technological change.

[76] See, for example, Andrew Grove (1996) on the market for computers and semiconductors and Clayton Christensen (1997) on the market for storage devices.

[77] Douglas Irwin and Peter Klenow (1994), Flamm (1996), pp. 305-424, and Elhanan Helpman and Manuel Trajtenberg (1998), pp. 111-119, present models of the semiconductor industry.

[78] The role of information technology in U.S. economic growth is discussed by the Economics and Statistics Administration (2000); comparisons among OECD countries are given by the Organisation for Economic Co-operation and Development (2000, 2003).

The stagflation of the 1970's greatly undermined the Keynesian Revolution, leading to a New Classical Counter-revolution led by Lucas (1981) that has transformed macroeconomics. The unanticipated American growth revival of the 1990's has similar potential for altering economic perspectives. In fact, this is already foreshadowed in a steady stream of excellent books on the economics of information technology[79]. We are the fortunate beneficiaries of a new agenda for economic research that will refresh our thinking and revitalize our discipline.

References

Abramovitz, Moses (1956), "Resources and Output Trends in the United States since 1870, American Economic Review, Vol. 46, No. 1, March, pp. 5-23.

_____ (1986), "Catching Up, Forging Ahead, and Falling Behind", Journal of Economic History, Vol. 46, No. 2, June, pp. 385-406.

Abramovitz, Moses, and Paul David (1999), "American Macroeconomic Growth in the Era of Knowledge-Based Progress: The Long-Run Perspective," in Robert E. Gallman and Stanley I. Engerman, eds., Cambridge Economic History of the United States, Cambridge, Cambridge University Press, pp. 1-92.

____ and _____ (2001), "Two Centuries of American Macroeconomic Growth from Exploitation of Resource Abundance to Knowledge-Driven Development," Stanford, Stanford Institute for Economic Policy Research, Policy Paper No. 01-005, August.

Acemoglu, Daron (2000), "Technical Change, Inequality, and the Labor Market," Journal of Economic Literature, Vol. 40, No. 1, March, pp. 7-72.

Aizcorbe, Ana, Stephen D. Oliner, and Daniel E. Sichel (2003), "Trends in Semiconductor Prices: Breaks and Explanations", Washington, Board of Governors of the Federal Reserve System, July.

[79] See, for example, Carl Shapiro and Hal Varian (1999), Brynjolfsson and Kahin (2000), and Choi and Whinston (2000).

Baily, Martin N. (2002), "The New Economy: Post Mortem or Second Wind?" Journal of Economic Perspectives, Vol. 16, No. 1, Winter, pp. 3-22.

Baily, Martin N., and Robert J. Gordon (1988), "The Productivity Slowdown, Measurement Issues, and the Explosion of Computer Power", Brookings Papers on Economic Activity, 2, pp. 347-420.

Baily, Martin N. and Robert Z. Lawrence (2001), "Do We Have a New E-conomy?" American Economic Review, Vol. 91, No. 2, May, pp. 308-313.

Baldwin, John R., and Tarek M. Harchaoui (2002), Productivity Growth in Canada – 2002, Ottawa, Statistics Canada.

Baumol, William J. (1986), "Productivity Growth, Convergence, and Welfare", American Economic Review, Vol. 76, No. 5, December, pp. 1072-1085.

Becker, Gary S. (1993a), Human Capital, 3rd ed., Chicago, University of Chicago Press (1st ed., 1964; 2nd ed., 1975).

_____ (1993b), "Nobel Lecture: The Economic Way of Looking at Behavior", Journal of Political Economy, Vol. 101, No. 3, June, pp. 385-409.

Berndt, Ernst R., and Jack Triplett (2000), eds., Fifty Years of Economic Measurement, Chicago, University of Chicago Press.

Blades, Derek (2001), Measuring Capital: A Manual on the Measurement of Capital Stocks, Consumption of Fixed Capital, and Capital Services, Paris, Organisation for Economic Co-operation and Development, April.

Bosworth, Barry P., and Jack Triplett (2000), "What's New About the New Economy? IT, Growth and Productivity", Washington, The Brookings Institution, October 20.

Brynjolfsson, Erik, and Lorin M. Hitt (2000), "Beyond Computation: Information Technology, Organizational Transformation and Business Performance," Journal of Economic Perspectives, Vol. 14, No. 4, Fall, pp. 23-48.

Brynjolfsson, Erik, and Brian Kahin (2000), eds., Understanding the Digital Economy, Cambridge, The MIT Press.

Brynjolfsson, Erik, and Shinkyo Yang (1996), "Information Technology and Productivity: A Review of the Literature", Advances in Computers, Vol. 43, No. 1, February, pp. 179-214.

Bureau of Economic Analysis (1986), "Improved Deflation of Purchase of Computers", Survey of Current Business, Vol. 66, No. 3, March, pp. 7-9.

_____ (1995), "Preview of the Comprehensive Revision of the National Income and Product Accounts: Recognition of Government Investment and Incorporation of a New Methodology for Calculating Depreciation," Survey of Current Business, Vol. 75, No. 9, September, pp. 33-41.

_____ (1999), Fixed Reproducible Tangible Wealth in the United States, 1925-94, Washington, U.S. Department of Commerce.

Bureau of Labor Statistics (1983), Trends in Multifactor Productivity, 1948-1981, Washington, U.S. Government Printing Office.

_____ (1993). "Labor Composition and U.S. Productivity Growth, 1948-90," Bureau of Labor Statistics Bulletin 2426, Washington, U.S. Department of Labor.

_____ (1994), "Multifactor Productivity Measures, 1991 and 1992," News Release USDL 94-327, July 11.

Campbell, John Y., and Robert J. Shiller (1998), "Valuation Ratios and the Long-run Stock Market Outlook", Journal of Portfolio Management, Vol. 24, No. 2, Winter, pp. 11-26.

Chamberlain, Gary (1982), "Multivariate Regression Models for Panel Data", Journal of Econometrics, Vol. 18, No. 1, January, pp. 5-46.

Chandler, Alfred D., Jr. (2000), "The Information Age in Historical Perspective," in Alfred D. Chandler and James W. Cortada, eds., A Nation Transformed by Information: How Information Has Shaped the United States from Colonial Times to the Present, New York, Oxford University Press, pp. 3-38.

Choi, Soon-Yong, and Andrew B. Whinston (2000), The Internet Economy: Technology and Practice, Austin, SmartEcon Publishing.

Chow, Gregory C. (1967), "Technological Change and the Demand for Computers," American Economic Review, Vol. 57, No. 5, December, pp. 1117-30.

Christensen, Clayton M. (1997), The Innovator's Dilemma, Boston, Harvard Business School Press.

Christensen, Laurits R., Dianne Cummings, and Dale W. Jorgenson (1980), "Economic Growth, 1947-1973: An International Comparison," in Kendrick and Vaccara, eds., pp. 595-698.

_____, _____ and _____ (1981), "Relative Productivity Levels", 1947-1973", European Economic Review, Vol. 16, No. 1, May, pp. 61-94.

Cole, Rosanne, Y.C.Chen, Joan A. Barquin-Stolleman, Ellen R. Dulberger, Nurthan Helvacian, and James H. Hodge (1986), "Quality-Adjusted Price Indexes for Computer Processors and Selected Peripheral Equipment," Survey of Current Business, Vol. 66, No. 1, January, pp. 41-50.

Colecchia, Alessandra, and Paul Schreyer (2002), "ICT Investment and Economic Growth in the 1990s: Is the United States a Unique Case? A Comparative Study of Nine OECD Countries", Review of Economic Dynamics, Vol. 5, No. 2, April 2002, pp. 408-442.

Congressional Budget Office (2002), The Budget and Economic Outlook: An Update, Washington, DC: U.S. Government Printing Office, July.

Corrado, Carol (2003), "Industrial Production and Capacity Utilization: The 2002 Historical and Annual Revision", Federal Reserve Bulletin, April, pp. 151-176.

Carol Corrado, John Haltiwanger, and Charles Hulten (2004), eds., Measurement of Capital in the New Economy, Chicago, University of Chicago Press, forthcoming.

Council of Economic Advisers (2002), Annual Report, Washington, U.S. Government Printing Office, February.

David, Paul A. (1990), "The Dynamo and the Computer: An Historical Perspective on the Productivity Paradox", American Economic Review, Vol. 80, No. 2, May, pp. 355-61.

_____ (2000), "Understanding Digital Technology's Evolution and the Path of Measured Productivity Growth: Present and Future in the Mirror of the Past", in Brynjolfsson and Kahin, eds., pp. 49-98.

Denison, Edward F. (1962), The Sources of Economic Growth in the United States and the Alternatives Before Us, New York, Committee on Economic Development.

_____ (1967), Why Growth Rates Differ, Washington, The Brookings Institution.

_____ (1989), Estimates of Productivity Change by Industry, Washington, Brookings Institution.

_____ (1993), "Robert J. Gordon's Concept of Capital," Review of Income and Wealth, Series 39, No. 1, March, pp. 89-102.

Dertouzos, Michael, Robert M. Solow and Richard K. Lester (1989), Made in American: Regaining the Productive Edge, Cambridge, The MIT Press.

Diewert, W. Erwin (1976), "Exact and Superlative Index Numbers", Journal of Econometrics, Vol. 4, No. 2, May, pp. 115-46.

_____ (1980), "Aggregation Problems in the Measurement of Capital, in Usher, ed., pp. 433-528.

Diewert, W. Erwin, and Denis A. Lawrence (2000), "Progress in Measuring the Price and Quantity of Capital," in Lau, ed., pp. 273-326.

Domar, Evsey (1946), "Capital Expansion, Rate of Growth and Employment", Econometrica, Vol. 14, No. 2, April, pp. 137-147.

_____ (1961), "On the Measurement of Technological Change," Economic Journal, Vol. 71, No. 284, December, pp. 709-29.

Doms, Mark (2004), "Communications Equipment: What Has Happened to Prices?" in Corrado, Haltiwanger, and Hulten, eds., forthcoming.

Dougherty, Chrys, and Dale W. Jorgenson (1996), "International Comparisons of the Sources of Economic Growth," American Economic Review, Vol. 86, No. 2, May, pp. 25-29.

_____ and _____ (1997), "There Is No Silver Bullet: Investment and Growth in the G7", National Institute Economic Review, No. 162, October, pp. 57-74.

Douglas, Paul H. (1948), "Are There Laws of Production?" American Economic Review, Vol. 38, No. 1, March, pp. 1-41.

Dulberger, Ellen R. (1989), "The Application of a Hedonic Model to a Quality-Adjusted Price Index for Computer Processors," in Jorgenson and Landau, eds., pp. 37-76.

_____ (1993), "Sources of Decline in Computer Processors: Selected Electronic Components", in Murray F. Foss, Marilyn E. Manser, and Allan H. Young, eds., Price Measurements and Their Uses, Chicago, University of Chicago Press, pp. 103-24.

Durlauf, Steven N., and Danny T. Quah (1999), "The New Empirics of Economic Growth", in Taylor and Woodford, eds., Vol. 1A, pp. 235-310.

Easterly, William (2001), The Elusive Quest for Growth, Cambridge, The MIT Press.

Economics and Statistics Administration (2000), Digital Economy 2000, Washington, DC: U.S. Department of Commerce, June.

Fisher, Irving (1922), The Making of Index Numbers, Boston, Houghton-Mifflin.

Flamm, Kenneth (1989), "Technological Advance and Costs: Computers versus Communications," in Robert C. Crandall and Kenneth Flamm, eds., Changing the Rules: Technological Change, International Competition, and Regulation in Communications, Washington, The Brookings Institution, pp. 13-61.

_____ (1996), Mismanaged Trade? Strategic Policy and the Semiconductor Industry, Washington, DC: Brookings Institution Press.

Fraumeni, Barbara M. (1997), "The Measurement of Depreciation in the U.S. National Income and Product Accounts", Survey of Current Business, Vol. 77, No. 7, July, pp. 7-23.

Goldsmith, Raymond, The National Wealth of the United States in the Postwar Period, New York, National Bureau of Economic Research.

Gollop, Frank M. (2000), "The Cost of Capital and the Measurement of Productivity," in Lau, ed., pp. 85-110.

Gordon, Robert J. (1989), "The Postwar Evolution of Computer Prices," in Jorgenson and Landau, pp. 77-126.

_____ (1990), The Measurement of Durable Goods Prices, Chicago, University of Chicago Press.

_____ (1998), "Foundations of the Goldilocks Economy: Supply Shocks and the Time-Varying NAIRU", Brookings Papers on Economic Activity, 2, pp. 297-333.

_____ (2000), "Does the 'New Economy' Measure Up to the Great Inventions of the Past," Journal of Economic Perspectives, Vol. 14, No. 4, Fall, pp. 49-74.

Greenspan, Alan, "Challenges for Monetary Policy-Makers," Washington, Board of Governors, Federal Reserve System, October 19, 2000.

Greenwood, Jeremy, Zvi Hercowitz, and Per Krusell (1997), "Long-run Implications of Investment-specific Technological Change," American Economic Review, Vol. 87, No. 3, June, pp. 342-62.

_____, _____ and _____ (2000), "The Role of Investment-specific Technological Change in the Business Cycle," European Economic Review, Vol. 44, No. 1, January, pp. 91-115.

Greenwood, Jeremy and Boyan Jovanovic (2001), "Accounting for Growth", in Hulten, Dean, and Harper, eds., pp. 179-222.

Griliches, Zvi (1960), "Measuring Inputs in Agriculture: A Critical Survey", Journal of Farm Economics, Vol. 40, No. 5, December, pp. 1398-1427.

_____ (1994), "Productivity, R&D, and the Data Constraint", American Economic Review, Vol. 94, No. 2, March, pp. 1-23.

_____ (2000), R&D, Education, and Productivity, Cambridge, Harvard University Press.

Grimm, Bruce T. (1997), "Quality Adjusted Price Indexes for Digital Telephone Switches," Washington, Bureau of Economic Analysis, May 20.

_____ (1998), "Price Indexes for Selected Semiconductors: 1974-96," Survey of Current Business, Vol. 78, No. 2, February, pp. 8-24.

Grove, Andrew S. (1996), Only the Paranoid Survive: How to Exploit the Crisis Points that Challenge Every Company, New York, Doubleday.

Gullickson, William, and Michael J. Harper (1999), "Possible Measurement Bias in Aggregate Productivity Growth," Monthly Labor Review, Vol. 122, No. 2, February, pp. 47-67.

Hall, Robert E. (1988), "The Relation between Price and Marginal cost in U.S. Industry", Journal of Political Economy, Vol. 96, No. 5, October, pp. 921-947.

_____ (1990a), "Invariance Properties of Solow's Productivity Residual", in Peter Diamond, ed., Growth/Productivity/Employment, Cambridge, The MIT Press.

_____ (2000) "e-Capital: The Link between the Stock Market and the Labor Market in the 1990's," Brookings Papers on Economic Activity, 2, pp. 73-118.

_____ (2001), "The Stock Market and Capital Accumulation", American Economic Review, Vol. 91, No. 5, December, pp. 1185-1202.

_____ (2002), Digital Dealing: How E-Markets Are Transforming the Economy, New York, Norton.

Harberger, Arnold C. (1998), "A Vision of the Growth Process", American Economic Review, Vol. 88, No. 1, March, pp. 1-32.

Harrod, Roy (1939), "An Essay in Dynamic Theory", Economic Journal, Vol. 49, No. 194, March, pp. 14-33.

Hayashi, Fumio (2000), "The Cost of Capital, Q, and the Theory of Investment Demand," in Lau, ed., pp. 55-84.

Hecht, Jeff (1999), City of Light, New York: Oxford University Press.

Helpman, Elhanan, and Manuel Trajtenberg (1998), "Diffusion of General Purpose Technologies," in Elhanan Helpman, ed., General Purpose Technologies and Economic Growth, Cambridge, The MIT Press, pp. 85-120.

Hercowitz, Zvi (1998), "The 'Embodiment' Controversy: A Review Essay," Journal of Monetary Economics, Vol. 41, No. 1, February, pp. 217-24.

Herman, Shelby W. (2000), "Fixed Assets and Consumer Durable Goods for 1925-99", Survey of Current Business, Vol. 80, No. 9, September, pp. 19-30.

Heston, Alan, Robert Summers, and Bettina Aten (2002), Penn World Table Version 6.1, Center for International Comparisons and the University of Pennsylvania (CICUP), October.

Hulten, Charles R. (1990), "The Measurement of Capital", in Berndt and Triplett, eds., pp. 119-152.

_____ (2001), "Total Factor Productivity: A Short Biography", in Hulten, Dean, and Harper, eds., pp. 1-47.

Hulten, Charles R., Edwin R. Dean, and Michael J. Harper (2001), eds., New Developments in Productivity Analysis, Chicago, University of Chicago Press.

International Technology Roadmap for Semiconductors, 2000 Update, Austin, Sematech Corporation, December 2000.

Irwin, Douglas A., and Peter J. Klenow (1994), "Learning-by-Doing Spillovers in the Semiconductor Industry," Journal of Political Economy, Vol. 102, No. 6, December, pp. 1200-27.

Islam, Nasrul (1995), "Growth Empirics", Quarterly Journal of Economics, Vol. 110, No. 4, November, pp. 1127-1170.

_____ (2003), "What Have We Learned from the Convergence Debate?" Journal of Economic Surveys, Vol. 17, No. 3, July, pp. 309-362.

Jorgenson, Dale W. (1963), "Capital Theory and Investment Behavior", American Economic Review, Vol. 53, No. 2, May, pp. 247-259.

_____ (1966), "The Embodiment Hypothesis", Journal of Political Economy, Vol. 74, No. 1, February, pp. 1-17.

_____ (1973), "The Economic Theory of Replacement and Depreciation", in Willy Sellekaerts, ed., Econometrics and Economic Theory, New York, Macmillan, pp. 189-221.

_____ (1990), "Productivity and Economic Growth", in Berndt and Triplett, eds., pp. 19-118.

_____ (1996), "Empirical Studies of Depreciation", Economic Inquiry, Vol. 34, No. 1, January, pp. 24-42.

_____ (2001), "Information Technology and the U.S. Economy", American Economic Review, Vol. 91, No. 1, March, 1-32.

Jorgenson, Dale W., Frank M. Gollop, and Barbara M. Fraumeni (1987), Productivity and U.S. Economic Growth, Cambridge, Harvard University Press.

Jorgenson, Dale W., and Zvi Griliches (1967), "The Explanation of Productivity Change," Review of Economic Studies, Vol. 34, No. 99, July, pp. 249-280.

Jorgenson, Dale W., Mun S. Ho, and Kevin J. Stiroh (2004), "Growth of U.S. Industries and Investments in Information Technology and Higher Education", in Carol Corrado, Charles Hulten, and Daniel Sichel, eds., Measuring Capital in a New Economy, Chicago, University of Chicago Press, forthcoming.

Jorgenson, Dale W., and Ralph Landau, eds., Technology and Capital Formation, Cambridge, The MIT Press, 1989.

Jorgenson, Dale W., and Kazuyuki Motohashi (2003), "Economic Growth of Japan and the U.S. in the Information Age", Tokyo, Research Institute of Economy, Trade, and Industry, July.

Jorgenson, Dale W., and Kevin J. Stiroh (1995), "Computers and Growth," Economics of Innovation and New Technology, Vol. 3, Nos. 3-4, pp. 295-316.

_____ and _____ (1999), "Information Technology and Growth," American Economic Review, Vol. 89, No. 2, May, pp. 109-15.

_____ and _____ (2000a), "U.S. Economic Growth at the Industry Level," American Economic Review, Vol. 90, No. 2, May, pp. 161-7.

_____ and _____ (2000b), "Raising the Speed Limit: U.S. Economic Growth in the Information Age," Brookings Papers on Economic Activity, 1, pp. 125-211.

Jorgenson, Dale W., and Eric Yip (2000), "Whatever Happened to Productivity Growth?" in Charles R. Hulten, Edwin R. Dean, and Michael J. Harper, eds., New Developments in Productivity Analysis, Chicago, University of Chicago Press, pp. 509-540.

Jorgenson, Dale W., and Kun-Young Yun (2001), Lifting the Burden: Tax Reform, the Cost of Capital, and U.S. Economic Growth, Cambridge, The MIT Press.

Katz, Lawrence F.)2000), "Technological Change, Computerization, and the Wage Structure," in Brynjolfsson and Kahin, eds., pp. 217-44.

Katz, Lawrence F., and Alan Krueger (1999), "The High-Pressure U.S. Labor Market of the 1990's," Brookings Papers on Economic Activity, 1, pp. 1-87.

Kendrick, John W. (1956), "Productivity Trends: Capital and Labor", Review of Economics and Statistics, Vol. 38, No. 3, August, pp. 248-257.

_____ (1961), Productivity Trends in the United States, Princeton, Princeton University Press.

_____ (1973), Postwar Productivity Trends in the United States, New York, National Bureau of Economic Research.

Kendrick, John W., and Eliot Grossman (1980), Productivity in the United States: Trends and Cycles, Baltimore, Johns Hopkins Press.

Kendrick, John W., and Beatrice Vaccara (1980), eds., New Developments in Productivity Measurement and Analysis, Chicago, University of Chicago Press.

Kiley, Michael T. (1999), "Computers and Growth with Costs of Adjustment: Will the Future Look Like the Past?" Washington, DC: Board of Governors of the Federal Reserve System, July.

King, Robert G., and Sergio Rebelo (1999), "Resuscitating Real Business Cycles", in Taylor and Woodford, eds., Vol. 1B, pp. 927-1008.

Konus, Alexander A., and S. S. Byushgens (1926), "On the Problem of the Purchasing Power of Money," Economic Bulletin of the Conjuncture Institute, Supplement, pp. 151-72.

Kravis, Irving B., Alan Heston, and Robert Summers (1978), International Comparisons of Real Product and Purchasing Power, Baltimore, Johns Hopkins University Press.

Kuznets, Simon (1971), Economic Growth of Nations, Cambridge, Harvard University Press.

Landefeld, J. Steven, and Robert P. Parker (1997), "BEA's Chain Indexes, Time Series, and Measures of Long-Term Growth," Survey of Current Business, Vol. 77, No. 5, May, pp. 58-68.

Lau, Lawrence J. (2000), ed., Econometrics and the Cost of Capital, Cambridge, MA: The MIT Press.

Lindbeck, Assar (1992), ed., Nobel Lectures in Economic Sciences, 1969-1980, River Edge, New Jersey, World Scientific Publishing Co.

Lucas, Robert E., Jr. (1967), "Adjustment Costs and the Theory of Supply," Journal of Political Economy, 75, No. 4, Part 1, August, pp. 321-34.

_____ (1981), Studies in Business-Cycle Theory, Cambridge, The MIT Press.

Maddison, Angus (1982), Phases of Capitalist Development, Oxford, Oxford University Press.

_____ (1987), "Growth and Slowdown in Advanced Capitalist Economies: Techniques of Quantitative Assessment", Journal of Economic Literature, Vol.25, No. 2, June, pp. 649-698.

_____ (1991), Dynamic Forces in Capitalist Development, Oxford, Oxford University Press.

_____ (1995), Monitoring the World Economy, Paris, Organisation for Economic Co-operation and Development.

_____ (2001), The World Economy: A Millenial Perspective, Paris, Organisation for Economic Co-operation and Development.

Maler, Karl-Goran (1992), ed., Nobel Lectures in Economic Sciences, 1981-1990, River Edge, New Jersey, World Scientifc Publishing Co.

Mankiw, N. Gregory, David Romer, and David Weil (1992), "A Contribution to the Empirics of Economic Growth", Quarterly Journal of Economics, Vol. 107, No. 2, May, pp. 407-437.

McGrattan, Ellen, and James Schmitz (1999), in Taylor and Woodford, eds., Vol. 1A, pp. 669-737.

Moore, Gordon E. (1965), "Cramming More Components onto Integrated Circuits," Electonics, Vol. 38, No. 8, April 19, pp. 114-7.

_____ (1996), "Intel -- Memories and the Microprocessor," Daedalus, 125, No. 2, Spring, pp. 55-80.

_____ (1997), "An Update on Moore's Law," Santa Clara, CA: Intel Corporation, September 30.

Moss, Milton (1973), ed., The Measurement of Economic and Social Performance, New York, Columbia University Press.

Moulton, Brent R. (2000), "Improved Estimates of the National Income and Product Accounts for 1929-99: Results of the Comprehensive Revision", Survey of Current Business, Vol. 80, No. 4, April, pp. 11-17, 36-145.

Oliner, Stephen D., and Daniel E. Sichel (1994), "Computers and Output Growth Revisited: How Big is the Puzzle?" Brookings Papers on Economic Activity, 2, pp. 273-317.

_____ and _____ (2000), "The Resurgence of Growth in the Late 1990's: Is Information Technology the Story?" Journal of Economic Perspectives, Vol. 14, No. 4, Fall, pp. 3-22.

Organisation for Economic Co-operation and Development (2000), A New Economy? Paris, Organisation for Economic Co-operation and Development.

_____ (2002), Purchasing Power Parities and Real Expenditures, 1999 Benchmark Year, Paris, Organization for Economic Co-operation and Development.

_____ (2003), ICT and Economic Growth, Paris, Organisation for Economic Co-operation and Development.

Parker, Robert P., and Bruce T. Grimm (2000), "Recognition of Business and Government Expenditures on Software as Investment: Methodology and Quantitative Impacts, 1959-98," Washington, Bureau of Economic Analysis, November 14.

Petzold, Charles (1999), Code: The Hidden Language of Computer Hardware and Software, Redmond, Microsoft Press.

Rashad, Rick (2000), "The Future -- It Isn't What It Used to Be," Seattle, Microsoft Research, May 3.

Romer, Paul (1987), "Crazy Explanations for the Productivity Slowdown", in Stanley Fischer, ed., NBER Macroeconomics Annual, Cambridge, The MIT Press, pp. 163-201.

_____ (1994), "The Origins of Endogenous Growth", Journal of Economic Perspectives, Vol. 8, No. 1, Winter, pp. 3-20.

Rees, Albert (1979), ed., Measurement and Interpretation of Productivity, Washington, National Academy Press.

Ruttan, Vernon W. (2001), "The Computer and Semiconductor Industries," in Technology, Growth, and Development, New York, Oxford University Press, pp. 316-67.

Samuelson, Paul A. (1961), "The Evaluation of 'Social Income': Capital Formation and Wealth", in Friedrich A. Lutz and Douglas C. Hague, eds., The Theory of Capital, London, Macmillan, pp. 32-57.

Schreyer, Paul (2000), "The Contribution of Information and Communication Technology to Output Growth: A Study of the G7 Countries", Paris, Organisation for Economic Co-operation and Development, May 23.

_____ (2001), OECD Productivity Manual: A Guide to the Measurement of Industry-Level and Aggregate Productivity Growth, Paris, Organisation for Economic Co-operation and Development, March.

Shapiro, Carl, and Hal R. Varian (1999), Information Rules, Boston, Harvard Business School Press.

Shiller, Robert (2000), Irrational Exuberance, Princeton, Princeton University Press.

Solow, Robert M. (1956), "A Contribution to the Theory of Economic Growth", Quarterly Journal of Economics, Vol. 70, No. 1, February, pp. 65-94.

_____ (1957), "Technical Change and the Aggregate Production Function", Review of Economics and Statistics, Vol. 39, No. 3, August, pp. 312-20.

_____ (1960), "Investment and Technical Progress", in Kenneth J. Arrow, Samuel Karlin, and Patrick Suppes, eds., Mathematical Methods in the Social Sciences, 1959, Stanford, Stanford University Press, pp. 89-104.

_____ (1970), *Growth Theory: An Exposition*, New York, Oxford University Press.

_____ (1987), "We'd Better Watch Out," New York Review of Books, July 12.

_____ (1988), "Growth Theory and After", American Economic Review, Vol. 78, No. 3, June, pp. 307-317.

_____ (1999), "Neoclassical Growth Theory," in Taylor and Woodford, eds., Vol. 1A, pp. 637-668.

_____ (2001), "After 'Technical Progress and the Aggregate Production Function'", in Hulten, Dean, and Harper, eds., pp. 173-178.

Stigler, George J. (1947), Trends in Output and Employment, New York: National Bureau of Economic Research.

Stiroh, Kevin J. (1998), "Computers, Productivity, and Input Substitution," Economic Inquiry, Vol. 36, No. 2, April, pp. 175-91.

_____ (2002), "Information Technology and the U.S. Productivity Revival: What Do the Industry Data Say?" American Economic Review, Vol. 92, No. 5, December, pp. 1559-76.

Summers, Robert, and Alan Heston (1984), "Improved International Comparisons of Real Product and Its Composition: 1950-1980," Review of Income and Wealth, Series 30, No. 1, March, pp. 1-25 (Mark 3).

_____ and _____ (1988), "A New Set of International Comparisons of Real Product and Price Levels: Estimates for 130 Countries, 1950-1985," Review of Income and Wealth, Series 34, No. 1, March, pp. 19-26 (Mark 4).

_____ and _____ (1991), "The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950-1988", Quarterly Journal of Economics, Vol. 106, No. 2, May, pp. 327-368.

Taylor, John B., and Michael Woodford (1999), eds., Handbook of Macroeconomics, 3 Vols., Amsterdam, North-Holland.

Tinbergen, Jan (1942), "On the Theory of Trend Movements", in Jan Tinbergen, Selected Papers, Amsterdam, North-Holland, 1959, pp. 182-221 (translated from "Zur Theorie der Langfristigen Wirtschaftsentwicklung", Weltwirtschaftliches Archiv, Band 55, Nu. 1, pp. 511-549).

Triplett, Jack E. (1986), "The Economic Interpretation of Hedonic Methods", Survey of Current Business, Vol. 66, No. 1, January, pp. 36-40.

_____ (1989), "Price and Technological Change in a Capital Good: Survey of Research on Computers," in Jorgenson and Landau, pp. 127-213.

_____ (1996), "High-tech Industry Productivity and Hedonic Price Indices," in Organisation for Economic Co-operation and Development, Industry Productivity, Paris, Organisation for Economic Co-operation and Development, 1996, pp. 119-42.

_____ (1999), "The Solow Productivity Paradox: What Do Computers Do to Productivity?" Canadian Journal of Economics, Vol. 32, No. 2, April, pp. 309-34.

_____ (2003), Handbook on the Quality Adjustment of Price Indices for ICT Products, Paris, Organisation for Economic Co-Operation and Development, forthcoming.

United Nations (1993), System of National Accounts 1993, New York, United Nations.

Usher, Daniel (1980), ed., The Measurement of Capital, Chicago, University of Chicago Press.

Van Ark, Bart, Johanna Melka, Nanno Mulder, Marcel Timmer, and Gerard Ypma (2002), ICT Investment and Growth Accounts for the European Union, 1980-2000, Brussels, European Commission, June.

Whelan, Karl (2002), "Computers, Obsolescence, and Productivity," Review of Economics and Statistics, Vol. 84, No. 3, August, pp. 445-62.

Wolfe, Tom (2000), "Two Men Who Went West", Hooking Up, New York, Farrar, Straus, and Giroux, pp. 17-65.

Wyckoff, Andrew W. (1995), "The Impact of Computer Prices on International Comparisons of Productivity," Economics of Innovation and New Technology, Vol. 3 Nos. 3-4, pp. 277-93.

Young, Allan (1989), "BEA's Measurement of Computer Output", Survey of Current Business, Vol. 69, No. 7, July, pp. 108-15.

Table 2.1: Information Technology Output and Gross Domestic Product

	Computer		Software		Communications		IT Services		Total IT		Gross Domestic Product	
Year	Value	Price	Value	Price	Value	Price	Value	Price	Value	Price	Value	Price
1948					1.7	0.90	0.4	6.52	2.1	4.53	321.0	0.19
1949					1.5	0.90	0.4	4.78	1.9	4.28	322.0	0.19
1950					1.7	0.92	0.5	4.51	2.3	4.30	343.4	0.19
1951					2.0	0.96	0.7	4.65	2.7	4.46	382.1	0.19
1952					2.5	0.93	1.0	5.13	3.6	4.47	395.1	0.19
1953					2.8	0.89	1.3	5.16	4.1	4.35	436.6	0.20
1954					2.5	0.90	1.1	3.69	3.6	3.94	428.2	0.19
1955					2.7	0.89	1.6	4.56	4.3	4.22	471.8	0.20
1956					3.4	0.91	1.8	4.60	5.2	4.27	493.7	0.21
1957					4.0	0.95	1.7	3.61	5.7	4.07	537.2	0.22
1958					3.5	0.95	1.9	3.73	5.4	4.13	512.6	0.21
1959	0.0	1,635.07			4.3	0.95	2.4	4.29	6.8	4.34	556.9	0.21
1960	0.2	1,635.07	0.1	1.25	4.8	0.93	2.5	4.02	7.6	4.18	573.1	0.22
1961	0.3	1,226.30	0.2	1.22	5.3	0.91	2.6	3.81	8.4	4.03	587.6	0.22
1962	0.3	817.53	0.2	1.18	5.8	0.91	3.0	4.04	9.3	4.03	631.3	0.22
1963	0.8	572.27	0.5	1.14	5.9	0.90	3.0	3.67	10.1	3.79	675.9	0.23
1964	1.0	490.52	0.6	1.11	6.5	0.88	3.2	3.57	11.3	3.67	737.6	0.23
1965	1.3	408.77	0.8	1.08	7.6	0.87	4.2	4.04	13.9	3.71	806.8	0.24
1966	1.9	283.63	1.1	0.99	9.1	0.85	4.6	3.76	16.7	3.43	881.7	0.25
1967	2.1	228.43	1.4	1.02	10.0	0.86	4.5	3.08	18.0	3.20	928.7	0.26
1968	2.1	194.16	1.5	1.01	10.7	0.87	4.9	2.88	19.2	3.12	981.9	0.27
1969	2.7	176.76	2.1	1.07	12.1	0.89	5.3	2.71	22.2	3.09	1,052.5	0.28
1970	3.0	158.80	2.8	1.14	13.4	0.92	6.0	2.75	25.2	3.13	1,111.2	0.29
1971	3.2	142.57	2.9	1.11	13.7	0.94	7.1	2.95	26.9	3.17	1,182.6	0.30
1972	4.1	128.28	3.4	1.10	14.4	0.96	8.0	3.01	29.9	3.18	1,323.8	0.32
1973	4.2	142.83	3.9	1.12	16.9	0.97	10.7	3.62	35.7	3.43	1,509.2	0.35
1974	4.8	128.86	4.8	1.18	18.4	1.02	9.7	2.97	37.7	3.29	1,628.7	0.38
1975	4.6	152.47	5.9	1.25	19.7	1.09	10.6	2.97	40.8	3.49	1,808.8	0.42
1976	5.6	125.12	6.4	1.25	22.0	1.12	12.5	3.20	46.5	3.54	2,054.9	0.46
1977	7.2	98.56	6.8	1.27	26.0	1.10	19.9	4.65	59.9	3.82	2,270.7	0.49
1978	9.7	60.47	8.0	1.27	30.3	1.13	17.9	3.78	65.9	3.41	2,547.6	0.51
1979	13.2	45.21	10.2	1.30	35.7	1.16	23.2	4.37	82.2	3.44	2,878.4	0.56
1980	17.3	34.17	12.3	1.35	40.7	1.22	20.8	3.50	91.0	3.17	3,011.1	0.59
1981	22.6	25.95	14.9	1.42	45.1	1.29	19.7	2.95	102.3	2.99	3,341.7	0.64
1982	25.4	25.83	17.7	1.45	46.9	1.34	22.5	2.99	112.5	3.05	3,532.2	0.69
1983	34.8	20.42	20.9	1.44	50.4	1.35	25.9	2.98	132.0	2.89	3,886.1	0.72
1984	43.4	18.70	25.9	1.43	57.8	1.36	30.4	2.97	157.4	2.82	4,375.0	0.76
1985	46.0	15.41	30.1	1.41	64.1	1.35	34.1	2.86	174.2	2.64	4,624.7	0.76
1986	45.7	13.64	32.7	1.36	57.9	1.37	38.2	2.73	174.6	2.54	4,753.7	0.75
1987	48.6	12.40	37.8	1.36	58.4	1.35	43.3	2.64	188.0	2.44	5,118.9	0.78
1988	54.0	12.15	44.7	1.35	63.9	1.32	51.6	2.71	214.3	2.42	5,702.9	0.84
1989	56.8	12.01	54.2	1.30	66.5	1.31	54.8	2.53	232.3	2.35	6,028.4	0.86
1990	52.3	10.86	62.3	1.26	69.5	1.31	59.5	2.45	243.6	2.27	6,339.7	0.89
1991	52.5	10.77	70.8	1.25	66.9	1.33	58.7	2.15	249.0	2.20	6,464.4	0.90
1992	55.3	9.76	76.7	1.16	70.5	1.31	66.9	2.17	269.4	2.10	6,795.1	0.92
1993	56.3	8.57	86.1	1.14	76.7	1.29	72.6	2.06	291.7	2.00	7,038.5	0.93
1994	60.4	8.19	93.4	1.11	84.3	1.26	83.7	2.06	321.8	1.96	7,579.5	0.96
1995	74.9	5.61	102.0	1.09	94.4	1.21	95.7	2.03	366.9	1.78	7,957.2	0.98
1996	84.8	3.53	115.4	1.05	107.8	1.18	103.6	1.83	411.5	1.55	8,475.4	1.00
1997	94.2	2.43	142.3	1.00	119.2	1.17	109.4	1.57	465.1	1.36	8,961.0	1.00
1998	96.6	1.69	162.5	0.97	124.1	1.11	127.1	1.46	510.2	1.22	9,346.9	0.99
1999	101.9	1.22	194.7	0.97	134.0	1.05	130.2	1.21	560.8	1.08	9,824.2	0.99
2000	109.9	1.00	222.7	1.00	152.6	1.00	130.8	1.00	616.0	1.00	10,399.6	1.00
2001	98.6	0.79	219.6	1.01	146.5	0.95	135.2	0.88	599.9	0.92	10,628.5	1.01
2002	88.4	0.64	212.8	1.00	127.4	0.90	136.7	0.77	565.2	0.85	11,279.4	1.04

Notes: Values are in billions of current dollars. Price are normalized to one in 2000. Information technology output is gross domestic product by type of product.

Table 2.2: Growth Rates of Outputs and Inputs

	1989-95		1995-2002	
	Prices	**Quantities**	**Prices**	**Quantities**
		Outputs		
Gross Domestic Product	2.11	2.52	0.96	4.02
Information Technology	-4.60	12.21	-10.55	16.72
Computers	-12.69	17.30	-30.99	33.37
Software	-2.82	13.34	-1.31	11.82
Communications Equipment	-1.36	7.19	-4.16	8.44
Information Technology Services	-3.66	12.95	-13.91	19.00
Non-Information Technology Investment	1.89	1.25	0.38	3.65
Non-Information Technology Consumption	2.52	2.34	1.93	3.22
		Inputs		
Gross Domestic Income	2.45	2.17	2.10	2.88
Information Technology Capital Services	-3.68	12.39	-10.49	18.11
Computer Capital Services	-10.28	19.99	-25.92	32.09
Software Capital Services	-4.20	14.76	-1.48	14.02
Communications Equipment Capital Services	0.99	5.99	-5.56	9.83
Non-Information Technology Capital Services	1.69	1.94	1.67	3.07
Labor Services	3.37	1.64	3.42	1.50

Notes: Average annual percentage rates of growth.

Table 2.3: Information Technology Capital Stock and Domestic Tangible Assets

	Computer		Software		Communications		Total IT		Total Domestic Tangible Assets	
Year	Value	Price	Value	Price	Value	Price	Value	Price	Value	Price
1948					4.6	0.93	4.6	1.99	754.9	0.11
1949					5.7	0.93	5.7	2.00	787.1	0.11
1950					7.0	0.95	7.0	2.04	863.5	0.11
1951					8.6	0.99	8.6	2.13	990.4	0.12
1952					10.0	0.96	10.0	2.05	1,066.5	0.12
1953					11.5	0.92	11.5	1.97	1,136.3	0.13
1954					12.9	0.93	12.9	1.99	1,187.7	0.13
1955					14.3	0.92	14.3	1.98	1,279.3	0.13
1956					16.4	0.94	16.4	2.01	1,417.8	0.14
1957					19.4	0.98	19.4	2.09	1,516.9	0.14
1958					21.1	0.98	21.1	2.11	1,586.0	0.14
1959	0.2	1,815.32	0.1	1.16	23.1	0.98	23.4	2.11	1,682.5	0.15
1960	0.2	1,773.96	0.1	1.16	24.9	0.96	25.2	2.06	1,780.8	0.15
1961	0.3	2,611.54	0.3	1.14	27.1	0.94	27.8	2.02	1,881.0	0.15
1962	0.5	2,188.01	0.4	1.10	29.9	0.94	30.8	2.00	2,007.2	0.16
1963	0.7	1,282.20	0.7	1.06	32.0	0.92	33.7	1.94	2,115.4	0.16
1964	0.8	738.09	1.0	1.04	34.5	0.91	37.1	1.90	2,201.2	0.16
1965	1.2	670.48	1.5	1.02	37.8	0.89	41.5	1.86	2,339.3	0.16
1966	2.2	665.29	2.1	0.94	42.1	0.88	47.1	1.78	2,534.9	0.17
1967	2.4	418.43	2.9	0.97	48.0	0.89	54.5	1.78	2,713.9	0.17
1968	2.6	324.89	3.4	0.96	54.4	0.91	62.1	1.79	3,004.5	0.18
1969	2.7	249.71	4.6	1.01	61.7	0.93	71.6	1.82	3,339.1	0.20
1970	3.6	242.63	6.2	1.09	70.0	0.96	82.5	1.87	3,617.5	0.21
1971	5.3	270.09	7.0	1.06	77.3	0.98	90.7	1.86	3,942.2	0.22
1972	4.9	179.70	8.1	1.05	85.2	1.01	100.7	1.87	4,463.6	0.24
1973	4.4	122.27	9.6	1.07	93.8	1.02	112.0	1.89	5,021.4	0.26
1974	6.6	143.74	11.7	1.12	105.8	1.07	126.7	1.94	5,442.4	0.27
1975	5.9	105.82	14.4	1.19	120.6	1.14	144.8	2.06	6,242.6	0.30
1976	6.6	96.27	16.3	1.19	133.0	1.18	159.8	2.09	6,795.1	0.32
1977	7.0	76.83	18.1	1.21	142.2	1.16	172.8	2.04	7,602.8	0.35
1978	11.8	83.34	20.4	1.21	160.3	1.19	194.9	2.03	8,701.7	0.38
1979	11.6	49.38	24.5	1.24	181.9	1.22	225.8	2.05	10,049.5	0.43
1980	16.6	43.74	29.6	1.29	210.5	1.28	264.4	2.09	11,426.5	0.47
1981	17.6	29.23	36.3	1.35	243.4	1.36	313.5	2.17	13,057.6	0.53
1982	19.6	22.05	43.2	1.39	270.6	1.40	356.4	2.20	14,020.9	0.55
1983	26.5	20.28	50.3	1.38	293.3	1.41	396.6	2.16	14,589.5	0.57
1984	36.2	18.16	60.1	1.37	320.6	1.42	447.3	2.11	15,901.1	0.60
1985	39.7	13.79	70.5	1.36	348.1	1.42	497.0	2.05	17,616.4	0.64
1986	43.3	11.29	79.3	1.31	374.2	1.40	540.3	1.96	18,912.3	0.67
1987	53.4	10.84	91.2	1.31	402.9	1.39	589.0	1.91	20,263.5	0.70
1988	52.6	8.55	105.4	1.30	432.9	1.37	646.5	1.87	21,932.4	0.74
1989	57.7	7.70	121.9	1.25	461.6	1.36	706.0	1.83	23,678.3	0.78
1990	65.0	7.46	140.6	1.22	487.5	1.35	751.7	1.77	24,399.0	0.79
1991	64.8	6.56	163.2	1.22	508.1	1.34	797.0	1.73	24,896.4	0.79
1992	72.1	6.16	175.0	1.12	528.8	1.32	833.5	1.64	25,218.3	0.79
1993	78.2	5.34	199.2	1.11	550.7	1.30	888.8	1.58	25,732.9	0.79
1994	82.3	4.42	218.2	1.08	578.0	1.28	951.8	1.52	26,404.3	0.79
1995	103.2	4.16	242.7	1.07	605.5	1.24	1,026.5	1.44	28,003.7	0.82
1996	130.9	3.73	269.7	1.04	637.6	1.20	1,099.8	1.34	29,246.9	0.83
1997	141.5	2.77	312.4	1.00	678.7	1.18	1,203.6	1.25	31,146.2	0.86
1998	159.6	2.13	360.6	0.97	704.3	1.11	1,292.2	1.13	33,888.6	0.91
1999	163.4	1.48	433.7	0.97	741.3	1.05	1,427.2	1.05	36,307.6	0.95
2000	153.2	1.00	515.5	1.00	805.2	1.00	1,609.0	1.00	39,597.1	1.00
2001	171.8	0.88	563.7	1.01	844.3	0.95	1,687.6	0.94	42,566.9	1.05
2002	158.9	0.68	583.9	1.00	874.0	0.91	1,739.7	0.89	45,892.0	1.11

Notes: Values are in billions of current dollars. Prices are normalized to one in 2000. Domestic tangible assets include fixed assets and consumer durable goods, land, and inventories.

Table 2.4: Information Technology Capital Services and Gross Domestic Income

	Computer		Software		Communications		Total IT		Gross Domestic Income	
Year	Value	Price	Value	Price	Value	Price	Value	Price	Value	Price
1948					1.7	1.21	1.7	8.26	321.0	0.14
1949					1.4	0.87	1.4	5.99	322.0	0.13
1950					1.7	0.86	1.7	5.92	343.4	0.14
1951					2.1	0.91	2.1	6.23	382.1	0.14
1952					2.6	0.97	2.6	6.64	395.1	0.14
1953					3.1	0.98	3.1	6.71	436.6	0.15
1954					2.5	0.70	2.5	4.78	428.2	0.15
1955					3.5	0.87	3.5	5.95	471.8	0.16
1956					4.0	0.89	4.0	6.11	493.7	0.16
1957					3.5	0.69	3.5	4.74	537.2	0.17
1958					3.9	0.70	3.9	4.78	512.6	0.16
1959	0.2	1,815.23	0.1	1.49	4.9	0.81	5.2	5.55	556.9	0.17
1960	0.2	1,773.88	0.1	1.46	5.1	0.76	5.3	5.22	573.1	0.17
1961	0.3	2,611.50	0.1	1.47	5.3	0.72	5.7	5.04	587.6	0.17
1962	0.5	2,187.99	0.2	1.54	6.3	0.77	7.0	5.30	631.3	0.18
1963	0.7	1,282.19	0.3	1.35	6.2	0.68	7.1	4.53	675.9	0.19
1964	0.8	738.08	0.4	1.27	6.8	0.69	8.0	4.30	737.6	0.20
1965	1.2	670.47	0.6	1.31	8.7	0.80	10.5	4.85	806.8	0.21
1966	2.2	665.29	1.0	1.39	9.2	0.75	12.4	4.63	881.7	0.22
1967	2.4	418.43	1.1	1.10	9.4	0.68	12.8	3.88	928.7	0.22
1968	2.6	324.89	1.5	1.23	9.8	0.64	14.0	3.59	981.9	0.22
1969	2.7	249.71	1.7	1.08	10.9	0.64	15.3	3.37	1,052.5	0.23
1970	3.6	242.62	2.3	1.13	12.7	0.68	18.5	3.51	1,111.2	0.24
1971	5.2	270.09	3.5	1.47	14.3	0.70	23.0	3.79	1,182.6	0.26
1972	4.9	179.70	3.7	1.32	16.0	0.73	24.5	3.51	1,323.8	0.28
1973	4.4	122.27	4.2	1.28	21.7	0.92	30.2	3.83	1,509.2	0.30
1974	6.6	143.74	4.9	1.29	19.5	0.76	30.9	3.48	1,628.7	0.32
1975	5.9	105.82	6.2	1.41	22.3	0.82	34.4	3.49	1,808.8	0.36
1976	6.6	96.27	7.0	1.39	23.9	0.82	37.5	3.44	2,054.9	0.39
1977	7.0	76.83	7.8	1.38	39.5	1.26	54.2	4.44	2,270.7	0.42
1978	11.8	83.34	9.0	1.45	33.6	0.98	54.4	3.83	2,547.6	0.45
1979	11.6	49.38	10.4	1.45	44.9	1.19	66.8	3.92	2,878.4	0.49
1980	16.6	43.74	12.2	1.46	40.0	0.96	68.8	3.34	3,011.1	0.51
1981	17.6	29.23	13.6	1.40	38.6	0.84	69.8	2.79	3,341.7	0.55
1982	19.6	22.05	15.2	1.34	41.4	0.83	76.2	2.55	3,532.2	0.58
1983	26.5	20.28	17.9	1.36	46.4	0.87	90.8	2.56	3,886.1	0.64
1984	36.2	18.16	22.2	1.42	53.9	0.93	112.3	2.58	4,375.0	0.68
1985	39.7	13.79	26.5	1.41	60.3	0.96	126.5	2.40	4,624.7	0.69
1986	43.3	11.29	30.8	1.40	67.3	0.98	141.3	2.27	4,753.7	0.70
1987	53.4	10.84	36.1	1.42	78.9	1.06	168.4	2.33	5,118.9	0.72
1988	52.6	8.55	44.2	1.50	97.2	1.20	194.0	2.35	5,702.9	0.77
1989	57.7	7.70	53.7	1.53	98.8	1.14	210.2	2.23	6,028.4	0.79
1990	65.0	7.46	59.3	1.42	102.5	1.10	226.8	2.14	6,339.7	0.81
1991	64.8	6.56	62.0	1.26	97.1	0.99	223.8	1.90	6,464.4	0.82
1992	72.1	6.16	81.6	1.42	105.4	1.02	259.1	1.96	6,795.1	0.85
1993	78.2	5.34	79.0	1.19	118.2	1.09	275.3	1.83	7,038.5	0.86
1994	82.3	4.42	96.0	1.27	132.6	1.14	310.9	1.81	7,579.5	0.90
1995	103.2	4.16	101.2	1.19	150.2	1.20	354.5	1.79	7,957.2	0.91
1996	130.9	3.73	114.6	1.19	144.2	1.07	389.6	1.65	8,475.4	0.94
1997	141.5	2.77	132.5	1.17	147.6	1.01	421.5	1.45	8,961.0	0.96
1998	159.6	2.13	150.1	1.10	184.4	1.15	494.1	1.37	9,346.9	0.96
1999	163.4	1.48	162.7	1.00	188.1	1.06	514.1	1.15	9,824.2	0.98
2000	153.2	1.00	190.3	1.00	201.4	1.00	544.9	1.00	10,399.6	1.00
2001	171.8	0.88	215.3	1.01	199.5	0.88	586.6	0.93	10,628.5	1.00
2002	127.8	0.56	243.3	1.07	202.5	0.82	604.6	0.86	11,279.4	1.06

Notes: Values are in billions of current dollars. Prices are normalized to one in 1996.

Table 2.5: Labor Services

	Labor Services					Weekly	Hourly	Hours
Year	Price	Quantity	Value	Quality	Employment	Hours	Compensation	Worked
1948	0.07	2,324.8	150.1	0.73	61,536	40.6	1.2	129,846
1949	0.07	2,262.8	165.5	0.73	60,437	40.2	1.3	126,384
1950	0.08	2,350.6	181.3	0.75	62,424	39.8	1.4	129,201
1951	0.08	2,531.5	210.7	0.76	66,169	39.7	1.5	136,433
1952	0.09	2,598.2	222.5	0.78	67,407	39.2	1.6	137,525
1953	0.09	2,653.0	238.5	0.79	68,471	38.8	1.7	138,134
1954	0.09	2,588.7	240.7	0.79	66,843	38.4	1.8	133,612
1955	0.09	2,675.7	252.7	0.80	68,367	38.7	1.8	137,594
1956	0.10	2,738.0	272.4	0.80	69,968	38.4	2.0	139,758
1957	0.11	2,740.9	293.0	0.81	70,262	37.9	2.1	138,543
1958	0.12	2,671.8	307.4	0.82	68,578	37.6	2.3	134,068
1959	0.12	2,762.8	316.9	0.82	70,149	37.8	2.3	137,800
1960	0.12	2,806.6	341.7	0.83	71,128	37.6	2.5	139,150
1961	0.12	2,843.4	352.1	0.84	71,183	37.4	2.5	138,493
1962	0.13	2,944.4	374.1	0.86	72,673	37.4	2.7	141,258
1963	0.13	2,982.3	382.7	0.86	73,413	37.3	2.7	142,414
1964	0.14	3,055.7	412.0	0.87	74,990	37.2	2.8	144,920
1965	0.14	3,149.7	448.1	0.87	77,239	37.2	3.0	149,378
1966	0.15	3,278.8	494.8	0.87	80,802	36.8	3.2	154,795
1967	0.16	3,327.2	518.9	0.88	82,645	36.3	3.3	156,016
1968	0.17	3,405.4	582.6	0.88	84,733	36.0	3.7	158,604
1969	0.18	3,491.1	641.4	0.88	87,071	35.9	4.0	162,414
1970	0.20	3,439.2	683.1	0.88	86,867	35.3	4.3	159,644
1971	0.22	3,439.5	740.7	0.89	86,715	35.3	4.7	158,943
1972	0.23	3,528.8	813.3	0.89	88,838	35.3	5.0	162,890
1973	0.25	3,672.4	903.9	0.89	92,542	35.2	5.3	169,329
1974	0.27	3,660.9	979.2	0.89	94,121	34.5	5.8	168,800
1975	0.29	3,606.4	1,055.2	0.90	92,575	34.2	6.4	164,460
1976	0.32	3,708.0	1,182.6	0.90	94,922	34.2	7.0	168,722
1977	0.35	3,829.8	1,321.1	0.90	98,202	34.1	7.6	174,265
1978	0.38	3,994.9	1,496.8	0.90	102,931	34.0	8.2	181,976
1979	0.40	4,122.6	1,660.4	0.90	106,463	33.9	8.9	187,589
1980	0.44	4,105.6	1,809.7	0.90	107,061	33.5	9.7	186,202
1981	0.47	4,147.7	1,934.6	0.91	108,050	33.3	10.4	186,887
1982	0.50	4,110.2	2,056.5	0.92	106,749	33.1	11.2	183,599
1983	0.54	4,172.3	2,234.7	0.92	107,810	33.2	12.0	186,175
1984	0.56	4,417.4	2,458.3	0.93	112,604	33.3	12.6	195,221
1985	0.58	4,531.7	2,646.2	0.93	115,201	33.3	13.3	199,424
1986	0.64	4,567.5	2,904.1	0.93	117,158	33.0	14.4	200,998
1987	0.64	4,736.5	3,017.3	0.94	120,456	33.1	14.6	207,119
1988	0.65	4,888.8	3,173.3	0.94	123,916	33.0	14.9	212,882
1989	0.68	5,051.3	3,452.4	0.95	126,743	33.2	15.8	218,811
1990	0.72	5,137.6	3,673.2	0.96	128,290	33.0	16.7	220,475
1991	0.75	5,086.7	3,806.3	0.96	127,022	32.7	17.6	216,281
1992	0.80	5,105.9	4,087.4	0.97	127,100	32.8	18.9	216,873
1993	0.82	5,267.6	4,323.8	0.97	129,556	32.9	19.5	221,699
1994	0.83	5,418.2	4,472.4	0.98	132,459	33.0	19.7	227,345
1995	0.84	5,573.2	4,661.5	0.98	135,297	33.1	20.0	232,675
1996	0.86	5,683.6	4,878.5	0.99	137,571	33.0	20.7	235,859
1997	0.89	5,843.3	5,186.5	0.99	140,432	33.2	21.4	242,242
1998	0.92	6,020.8	5,519.5	0.99	143,557	33.3	22.2	248,610
1999	0.96	6,152.1	5,908.2	1.00	146,468	33.3	23.3	253,276
2000	1.00	6,268.5	6,268.5	1.00	149,364	33.1	24.4	257,048
2001	1.05	6,250.6	6,537.4	1.01	149,020	32.9	25.6	255,054
2002	1.06	6,188.7	6,576.3	1.01	147,721	32.9	26.0	252,399

Notes: Value is in billions of current dollars. Quantity is in billions of 1996 dollars. Price and quality are normalized to one in 1996. Employment is in thousands of workers. Weekly hours is hours per worker, divided by 52. Hourly compensation is in current dollars. Hours worked are in millions of hours.

Table 2.6: Sources of Gross Domestic Product Growth

	1948-02	1948-73	1973-89	1989-95	1995-02
			Outputs		
Gross Domestic Product	3.46	3.78	3.06	2.52	4.02
Contribution of Information Technology	0.38	0.18	0.43	0.50	0.88
Computers	0.13	0.03	0.18	0.15	0.34
Software	0.07	0.02	0.08	0.15	0.19
Communications Equipment	0.08	0.07	0.08	0.08	0.11
Information Technology Services	0.10	0.05	0.09	0.13	0.24
Contribution of Non-Information Technology	3.08	3.61	2.63	2.02	3.14
Contribution of Non-Information Technology Investment	0.68	0.91	0.48	0.24	0.74
Contribution of Non-Information Technology Consumption	2.40	2.70	2.15	1.78	2.40
			Inputs		
Gross Domestic Income	2.79	2.99	2.68	2.17	2.88
Contribution of Information Technology Capital Services	0.36	0.15	0.38	0.48	0.91
Computers	0.17	0.04	0.20	0.22	0.50
Software	0.08	0.02	0.07	0.16	0.23
Communications Equipment	0.11	0.09	0.11	0.10	0.18
Contribution of Non-Information Technology Capital Services	1.39	1.79	1.15	0.72	1.09
Contribution of Labor Services	1.05	1.04	1.15	0.98	0.88
Total Factor Productivity	0.67	0.80	0.38	0.35	1.14

Notes: Average annual percentage rates of growth. The contribution of an output or input is the rate of growth, multiplied by the value share.

Table 2.7: Sources of Average Labor Productivity Growth

	1948-02	1948-73	1973-89	1989-95	1995-02
Gross Domestic Product	3.46	3.78	3.06	2.52	4.02
Hours Worked	1.23	1.06	1.60	1.02	1.16
Average Labor Productivity	2.23	2.72	1.46	1.50	2.86
Contribution of Capital Deepening	1.23	1.49	0.85	0.78	1.52
Information Technology	0.32	0.14	0.34	0.43	0.86
Non-Information Technology	0.90	1.35	0.51	0.35	0.66
Contribution of Labor Quality	0.33	0.43	0.23	0.36	0.20
Total Factor Productivity	0.67	0.80	0.38	0.35	1.14
Information Technology	0.17	0.05	0.20	0.23	0.47
Non-Information Technology	0.50	0.75	0.18	0.12	0.67
			Addendum		
Labor Input	1.81	1.83	1.99	1.64	1.50
Labor Quality	0.58	0.77	0.39	0.61	0.33
Capital Input	4.13	4.49	3.67	2.92	4.92
Capital Stock	3.29	4.13	2.77	1.93	2.66
Capital Quality	0.84	0.36	0.90	0.99	2.27

Notes: Average annual percentage rates of growth. Contributions are defined in Equation (3) of the text.

Table 2.8: Sources of Total Factor Productivity Growth

	1948-02	1948-73	1973-89	1989-95	1995-02
Total Factor Productivity Growth	0.67	0.80	0.38	0.35	1.14
		Contributions to TFP Growth:			
Information Technology	0.17	0.05	0.20	0.23	0.47
Computers	0.10	0.02	0.13	0.13	0.33
Software	0.02	0.00	0.03	0.06	0.06
Communications Equipment	0.04	0.03	0.05	0.04	0.08
Non-Information Technology	0.50	0.75	0.18	0.12	0.67
		Relative Price Changes:			
Information Technology	-6.72	-4.1	-8.5	-7.4	-11.7
Computers	-22.50	-22.0	-21.5	-15.1	-33.1
Software	-4.87	-5.1	-5.1	-5.3	-3.4
Communications Equipment	-3.79	-2.9	-4.1	-3.8	-6.3
Non-Information Technology	-0.45	-0.7	-0.1	-0.1	-0.5
		Average Nominal Shares:			
Information Technology	2.03	1.00	2.35	3.04	4.10
Computers	0.46	0.10	0.64	0.83	1.00
Software	0.53	0.07	0.49	1.13	1.78
Communications Equipment	1.04	0.83	1.22	1.09	1.33
Non-Information Technology	97.29	98.60	96.93	95.95	94.63

Notes: Average annual rates of growth. Prices are relative to the price of gross domestic income. Contributions are relative price changes, weighted by average nominal output shares.

Table 3.1 Levels of Output and Input Per Capita and Productivity

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				Output Per Capita			
1980	61.0	67.6	45.0	45.9	49.3	45.9	39.6
1989	76.9	78.8	56.5	54.1	58.6	57.3	56.0
1995	83.1	79.6	61.4	57.0	65.0	62.1	64.0
2001	99.9	91.8	71.3	64.0	69.2	68.8	70.6
				Input Per Capita			
1980	70.5	64.2	50.2	46.5	61.0	43.1	57.7
1989	83.9	74.4	61.2	53.3	71.1	55.5	72.0
1995	88.8	75.2	67.0	57.0	73.7	58.8	77.8
2001	100.8	83.7	73.6	61.7	79.0	67.2	80.9
				Productivity			
1980	86.6	105.4	89.5	98.6	80.8	106.6	68.7
1989	91.7	105.9	92.3	101.5	82.4	103.2	77.7
1995	93.6	105.9	91.7	99.9	88.1	105.6	82.3
2001	99.1	109.7	96.9	103.6	87.6	102.5	87.2

Note: U.S. = 100.0 in 2000, Canada data begins in 1981

Table 3.2 Growth Rate and Level in Output

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				Growth Rate (percentage)			
1980-1989	3.49	3.10	2.69	2.38	1.99	2.51	4.42
1989-1995	2.52	1.39	1.62	1.30	2.34	1.52	2.56
1995-2001	4.18	3.34	2.74	2.34	1.18	1.90	1.85
				Level (billions of 2000 U.S. Dollars)			
1980	5123.6	618.4	934.0	932.0	1421.7	955.7	1706.3
1989	7015.7	792.6	1190.3	1154.3	1700.2	1197.4	2539.3
1995	8161.2	861.4	1311.8	1247.8	1956.3	1311.5	2961.1
2001	10485.7	1052.3	1545.9	1436.0	2099.8	1470.1	3309.2
				Level (U.S. = 100.0 in 2000)			
1980	49.3	5.9	9.0	9.0	13.7	9.2	16.4
1989	67.5	7.6	11.4	11.1	16.3	11.5	24.4
1995	78.5	8.3	12.6	12.0	18.8	12.6	28.5
2001	100.8	10.1	14.9	13.8	20.2	14.1	31.8

Note: Canada data begins in 1981

Table 3.3 Growth Rate and Level in Population

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				Growth Rate			
1980-1989	0.92	1.18	0.16	0.54	0.05	0.05	0.59
1989-1995	1.23	1.22	0.24	0.45	0.62	0.18	0.33
1995-2001	1.12	0.95	0.24	0.41	0.14	0.18	0.22
				Level (millions)			
1980	227.7	24.8	56.3	55.1	78.3	56.4	116.8
1989	247.4	27.3	57.1	57.9	78.7	56.7	123.1
1995	266.3	29.4	58.0	59.4	81.7	57.3	125.6
2001	284.8	31.1	58.8	60.9	82.3	57.9	127.2
				Level (U.S. = 100.0 in 2000)			
1980	80.7	8.8	20.0	19.5	27.8	20.0	41.4
1989	87.7	9.7	20.3	20.5	27.9	20.1	43.6
1995	94.4	10.4	20.5	21.1	28.9	20.3	44.5
2001	101.0	11.0	20.8	21.6	29.2	20.5	45.1

Note: Percentage, Canada data begins in 1981

Table 3.4 Growth in Output and Input Per Capita and Productivity

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				Output per capita			
1980-1989	2.57	1.92	2.54	1.84	1.93	2.46	3.83
1989-1995	1.29	0.17	1.38	0.85	1.72	1.33	2.23
1995-2001	3.06	2.38	2.50	1.93	1.04	1.72	1.64
				Input Per Capita			
1980-1989	1.94	1.86	2.20	1.52	1.71	2.82	2.46
1989-1995	0.94	0.17	1.49	1.11	0.60	0.96	1.29
1995-2001	2.10	1.80	1.59	1.33	1.14	2.21	0.66
				Productivity			
1980-1989	0.63	0.06	0.34	0.32	0.23	-0.36	1.37
1989-1995	0.35	0.00	-0.11	-0.26	1.12	0.37	0.94
1995-2001	0.95	0.58	0.91	0.60	-0.10	-0.49	0.98

Note: Percentage, Canada data begins in 1981

Table 3.5 Levels of Capital Input and Capital Stock per capita and capital quality

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				Capital Input Per Capita			
1980	57.7	56.0	25.8	36.3	44.6	35.6	29.8
1989	73.7	67.1	37.9	48.3	62.1	62.4	42.1
1995	81.6	68.3	50.0	52.7	72.3	73.1	50.8
2001	103.9	78.0	56.1	58.1	83.5	89.4	58.9
				Capital Stock Per Capita			
1980	76.8	42.3	24.1	36.2	60.2	36.0	77.0
1989	88.4	47.9	31.2	42.4	67.9	52.4	82.8
1995	92.2	49.1	35.9	47.0	77.0	62.3	88.3
2001	101.7	55.1	44.5	52.0	85.5	72.3	93.5
				Capital Quality			
1980	75.1	132.3	107.0	100.1	74.0	98.8	38.6
1989	83.4	139.9	121.7	114.0	91.5	119.1	50.8
1995	88.5	139.1	139.3	112.2	94.0	117.4	57.5
2001	102.2	141.5	126.1	111.9	97.7	123.6	63.0

Note: U.S. = 100.0 in 2000, Canada data begins in 1981

Table 3.6 Levels of IT Capital Input and Capital Stock per capita and capital quality

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				IT Capital Input Per Capita			
1980	4.7	1.0	3.0	4.2	7.2	6.8	0.7
1989	19.7	4.0	11.0	12.0	18.9	19.0	5.6
1995	38.5	11.3	21.2	19.3	31.5	31.6	11.2
2001	115.2	46.2	54.3	38.7	60.5	61.1	39.7
				IT Capital Stock Per Capita			
1980	9.8	5.5	2.5	3.5	6.1	4.6	3.6
1989	27.4	10.3	9.6	9.9	15.5	13.1	11.2
1995	46.8	14.4	19.2	18.0	28.2	23.8	19.9
2001	110.7	21.6	44.9	33.4	49.7	44.1	71.0
				Capital Quality			
1980	48.0	17.6	120.1	119.1	119.0	148.8	20.1
1989	72.0	38.7	114.2	121.2	122.0	145.2	50.0
1995	82.3	78.9	110.3	107.6	111.5	132.7	56.5
2001	104.0	213.6	120.9	115.6	121.8	138.4	56.0

Note: U.S. = 100.0 in 2000, Canada data begins in 1981

Table 3.7 Levels of Non-IT Capital Input and Capital Stock per capita and capital quality

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				Non-IT Capital Input Per Capita			
1980	73.4	73.0	30.6	41.2	51.8	41.6	39.1
1989	86.6	82.9	43.3	53.8	70.1	71.1	51.2
1995	90.4	79.7	55.8	57.8	79.5	81.0	59.7
2001	102.3	83.9	56.3	62.5	87.1	94.5	60.8
				Non-IT Capital Stock Per Capita			
1980	82.5	44.4	25.7	38.0	63.4	38.2	82.8
1989	92.5	49.8	32.6	44.0	70.6	54.8	88.0
1995	94.8	50.7	36.9	48.3	79.3	64.4	93.1
2001	101.4	57.4	44.5	54.1	87.2	75.1	89.6
				Capital Quality			
1980	89.0	164.2	119.0	108.3	81.8	108.9	47.2
1989	93.7	166.4	132.9	122.4	99.3	129.7	58.2
1995	95.4	157.2	151.2	119.6	100.3	125.8	64.2
2001	100.9	146.1	126.5	115.6	99.9	125.8	67.8

Note: U.S. = 100.0 in 2000, Canada data begins in 1981

Table 3.8 Growth in Capital Input and Capital Stock Per Capita and Capital Quality

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				Capital Input Per Capita			
1980-1989	2.72	2.26	4.28	3.19	3.70	6.25	3.86
1989-1995	1.70	0.31	4.61	1.46	2.53	2.63	3.13
1995-2001	4.03	2.20	1.92	1.63	2.40	3.35	2.46
				Capital Stock Per Capita			
1980-1989	1.56	1.57	2.85	1.74	1.34	4.18	0.81
1989-1995	0.70	0.60	2.36	1.74	2.09	2.87	1.06
1995-2001	1.63	1.91	3.57	1.67	1.75	2.49	0.95
				Capital Quality			
1980-1989	1.17	0.69	1.43	1.45	2.36	2.07	3.05
1989-1995	0.99	-0.29	2.25	-0.27	0.44	-0.24	2.07
1995-2001	2.40	0.29	-1.65	-0.04	0.65	0.86	1.51

Note: Percentage, Canada data begins in 1981

Table 3.9 Growth in IT Capital Input and Capital Stock Per Capita and Capital Quality

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				IT Capital Input Per Capita			
1980-1989	15.98	17.66	14.43	11.66	10.71	11.44	22.74
1989-1995	11.16	17.42	10.91	7.92	8.47	8.44	11.57
1995-2001	18.24	23.42	15.69	11.55	10.87	10.98	21.08
				IT Capital Stock Per Capita			
1980-1989	11.47	7.83	14.98	11.46	10.43	11.72	12.61
1989-1995	8.94	5.53	11.50	9.91	9.97	9.94	9.52
1995-2001	14.34	6.82	14.16	10.35	9.40	10.28	21.22
				Capital Quality			
1980-1989	4.51	9.83	-0.56	0.20	0.28	-0.27	10.13
1989-1995	2.22	11.89	-0.58	-1.99	-1.50	-1.49	2.05
1995-2001	3.89	16.60	1.53	1.20	1.47	0.70	-0.14

Note: Percentage, Canada data begins in 1981

Table 3.10 Growth in Non-IT Capital Input and Capital Stock Per Capita and Capital Quality

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				Non-IT Capital Input Per Capita			
1980-1989	1.84	1.60	3.85	2.97	3.36	5.97	3.00
1989-1995	0.71	-0.66	4.22	1.20	2.09	2.17	2.58
1995-2001	2.05	0.85	0.15	1.30	1.52	2.57	0.29
				Non-IT Capital Stock Per Capita			
1980-1989	1.27	1.43	2.62	1.61	1.20	4.03	0.68
1989-1995	0.41	0.29	2.07	1.58	1.92	2.68	0.94
1995-2001	1.11	2.07	3.12	1.87	1.59	2.56	-0.63
				Capital Quality			
1980-1989	0.57	0.17	1.23	1.36	2.16	1.94	2.32
1989-1995	0.30	-0.95	2.15	-0.38	0.17	-0.51	1.64
1995-2001	0.94	-1.22	-2.97	-0.57	-0.06	0.01	0.92

Note: Percentage, Canada data begins in 1981

Table 3.11 Levels of Labor Input and Hours Worked Per Capita and Labor Quality

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				Labor Input Per Capita			
1980	81.1	73.0	78.9	63.0	75.4	48.8	91.4
1989	91.9	82.1	85.4	59.4	78.7	51.0	104.3
1995	94.2	82.3	82.4	61.7	75.2	50.6	103.9
2001	98.8	89.3	89.2	65.3	75.9	55.1	100.3
				Hours Worked Per Capita			
1980	89.7	91.4	92.0	79.3	82.3	71.4	116.9
1989	97.1	96.6	97.7	71.2	82.7	72.1	116.7
1995	95.9	90.9	89.8	67.6	76.4	68.9	109.9
2001	98.3	96.3	94.2	69.7	75.3	72.3	103.8
				Labor Quality			
1980	90.4	79.9	85.7	79.5	91.6	68.3	78.2
1989	94.7	85.0	87.4	83.5	95.2	70.7	89.4
1995	98.2	90.6	91.7	91.2	98.4	73.5	94.5
2001	100.5	92.7	94.7	93.7	100.9	76.1	96.6

Note: U.S. = 100.0 in 2000, Canada data begins in 1981

Table 3.12 Growth in Labor Input and Hours Worked Per Capita and Labor Quality

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				Labor Input Per Capita			
1980-1989	1.38	1.47	0.88	-0.65	0.48	0.49	1.47
1989-1995	0.41	0.04	-0.59	0.61	-0.78	-0.13	-0.07
1995-2001	0.79	1.35	1.32	0.95	0.17	1.40	-0.58
				Hours Worked Per Capita			
1980-1989	0.87	0.69	0.67	-1.20	0.06	0.10	-0.02
1989-1995	-0.21	-1.02	-1.41	-0.86	-1.33	-0.75	-0.99
1995-2001	0.41	0.98	0.79	0.50	-0.25	0.81	-0.95
				Labor Quality			
1980-1989	0.51	0.78	0.21	0.55	0.42	0.39	1.49
1989-1995	0.61	1.06	0.81	1.47	0.55	0.63	0.92
1995-2001	0.38	0.38	0.53	0.45	0.41	0.60	0.36

Note: Percentage, Canada data begins in 1981

Table 3.13 Contribution of Total Capital, IT Capital and Non-IT Capital to Output Growth

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				Total Capital			
1980-1989	1.53	1.71	1.80	2.12	1.44	2.55	1.85
1989-1995	1.19	0.76	1.96	1.12	1.31	1.12	1.47
1995-2001	2.10	1.67	0.94	1.15	1.11	1.47	1.10
				IT Capital			
1980-1989	0.45	0.39	0.24	0.18	0.19	0.24	0.43
1989-1995	0.48	0.49	0.27	0.19	0.26	0.26	0.31
1995-2001	0.97	0.86	0.76	0.42	0.46	0.49	0.75
				Non-IT Capital			
1980-1989	1.08	1.32	1.56	1.94	1.25	2.31	1.42
1989-1995	0.71	0.27	1.69	0.93	1.05	0.86	1.16
1995-2001	1.13	0.81	0.18	0.73	0.65	0.98	0.35

Note: Percentage. Contribution is growth rate times value share. Canada data begins in 1981

Table 3.14 Contributions of Productivity from IT and Non-IT Production to Output Growth

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				Productivity			
1980-1989	0.63	0.06	0.34	0.32	0.23	-0.36	1.37
1989-1995	0.35	0.00	-0.11	-0.26	1.12	0.37	0.94
1995-2001	0.95	0.58	0.91	0.60	-0.10	-0.49	0.98
				Productivity from IT Production			
1980-1989	0.23	0.14	0.23	0.29	0.28	0.32	0.23
1989-1995	0.23	0.14	0.32	0.29	0.43	0.38	0.29
1995-2001	0.48	0.17	0.82	0.56	0.65	0.68	0.57
				Productivity from Non-IT Production			
1980-1989	0.40	-0.08	0.11	0.03	-0.05	-0.68	1.14
1989-1995	0.12	-0.14	-0.43	-0.55	0.69	-0.01	0.65
1995-2001	0.47	0.41	0.09	0.04	-0.75	-1.17	0.41

Note: Percentage. Canada data begins in 1981

Table 3.15 Sources of Output Growth

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				Output			
1980-1989	3.49	3.10	2.69	2.38	1.99	2.51	4.42
1989-1995	2.52	1.39	1.62	1.30	2.34	1.52	2.56
1995-2001	4.18	3.34	2.74	2.34	1.18	1.90	1.85
				Labor			
1980-1989	1.33	1.33	0.56	-0.06	0.32	0.32	1.20
1989-1995	0.98	0.62	-0.24	0.44	-0.09	0.03	0.15
1995-2001	1.12	1.08	0.88	0.59	0.17	0.93	-0.22
				IT Capital			
1980-1989	0.45	0.39	0.24	0.18	0.19	0.24	0.43
1989-1995	0.48	0.49	0.27	0.19	0.26	0.26	0.31
1995-2001	0.97	0.86	0.76	0.42	0.46	0.49	0.75
				Non-IT Capital			
1980-1989	1.08	1.32	1.56	1.94	1.25	2.31	1.42
1989-1995	0.71	0.27	1.69	0.93	1.05	0.86	1.16
1995-2001	1.13	0.81	0.18	0.73	0.65	0.98	0.35
				Productivity from IT Production			
1980-1989	0.23	0.14	0.23	0.29	0.28	0.32	0.23
1989-1995	0.23	0.14	0.32	0.29	0.43	0.38	0.29
1995-2001	0.48	0.17	0.82	0.56	0.65	0.68	0.57
				Productivity from Non-IT Production			
1980-1989	0.40	-0.08	0.11	0.03	-0.05	-0.68	1.14
1989-1995	0.12	-0.14	-0.43	-0.55	0.69	-0.01	0.65
1995-2001	0.47	0.41	0.09	0.04	-0.75	-1.17	0.41

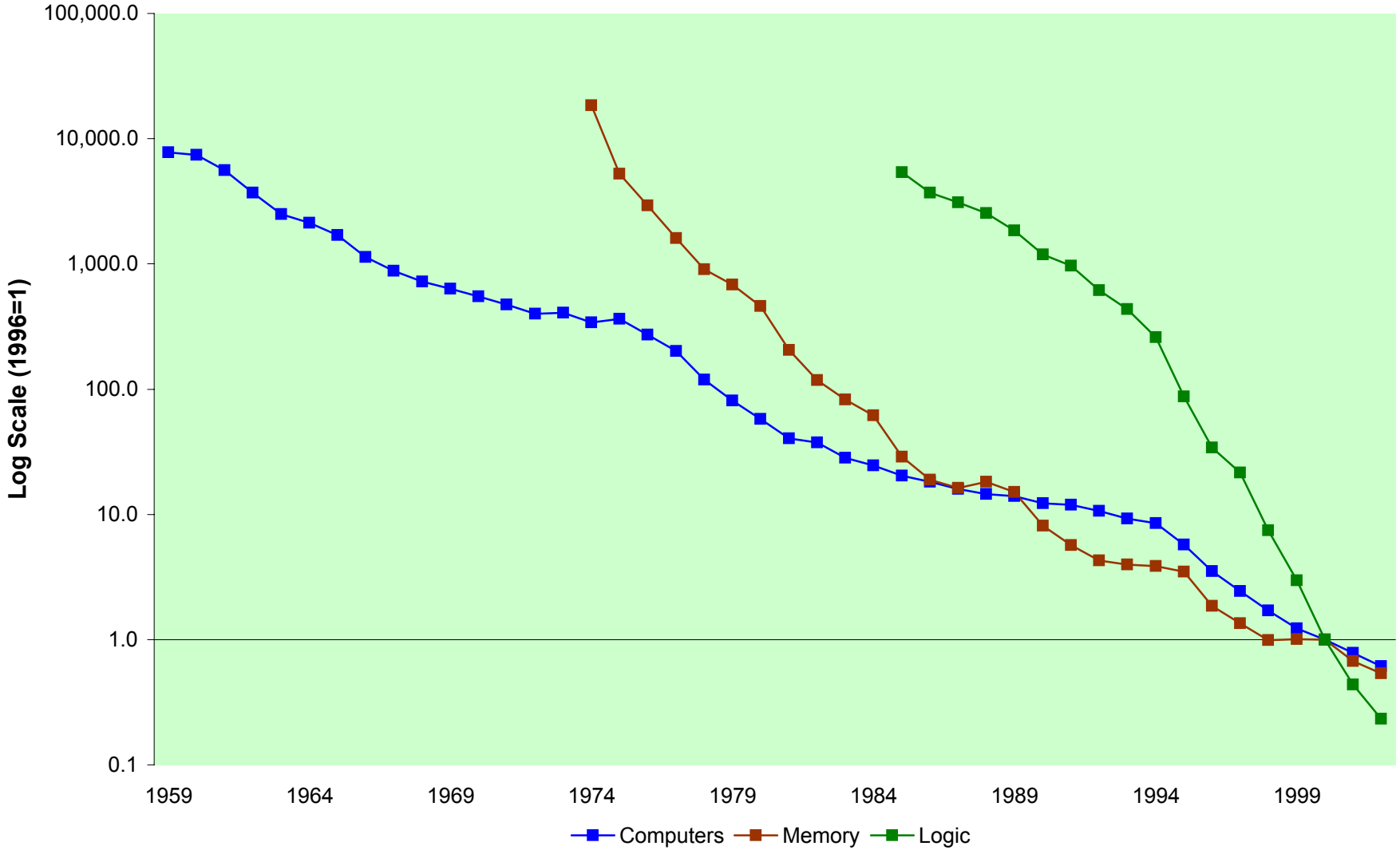
Note: Percentage. Contributions. Canada data begins in 1981

Table 3.16 Sources of Labor Productivity Growth

Year	U.S.	Canada	U.K.	France	Germany	Italy	Japan
				Output			
1980-1989	3.49	3.10	2.69	2.38	1.99	2.51	4.42
1989-1995	2.52	1.39	1.62	1.30	2.34	1.52	2.56
1995-2001	4.18	3.34	2.74	2.34	1.18	1.90	1.85
				Hours			
1980-1989	1.79	1.87	0.82	-0.66	0.11	0.15	0.56
1989-1995	1.02	0.20	-1.17	-0.41	-0.71	-0.57	-0.67
1995-2001	1.53	1.93	1.03	0.91	-0.11	0.99	-0.73
				Labor Productivity			
1980-1989	1.70	1.23	1.87	3.04	1.88	2.36	3.86
1989-1995	1.50	1.19	2.79	1.71	3.05	2.09	3.23
1995-2001	2.65	1.41	1.71	1.43	1.29	0.92	2.58
				IT Capital Deepening			
1980-1989	0.40	0.35	0.22	0.19	0.19	0.23	0.42
1989-1995	0.43	0.48	0.29	0.20	0.28	0.28	0.33
1995-2001	0.89	0.79	0.71	0.39	0.46	0.45	0.78
				Non-IT Capital Deepening			
1980-1989	0.37	0.42	1.20	2.29	1.20	2.25	1.20
1989-1995	0.35	0.16	2.11	1.15	1.33	1.06	1.42
1995-2001	0.58	-0.14	-0.21	0.25	0.70	0.61	0.61
				Labor Quality			
1980-1989	0.30	0.40	0.12	0.24	0.26	0.23	0.87
1989-1995	0.36	0.55	0.49	0.61	0.33	0.38	0.54
1995-2001	0.23	0.18	0.30	0.19	0.23	0.35	0.21
				Productivity from IT Production			
1980-1989	0.23	0.14	0.23	0.29	0.28	0.32	0.23
1989-1995	0.23	0.14	0.32	0.29	0.43	0.38	0.29
1995-2001	0.48	0.17	0.82	0.56	0.65	0.68	0.57
				Productivity from Non-IT Production			
1980-1989	0.40	-0.08	0.11	0.03	-0.05	-0.68	1.14
1989-1995	0.12	-0.14	-0.43	-0.55	0.69	-0.01	0.65
1995-2001	0.47	0.41	0.09	0.04	-0.75	-1.17	0.41

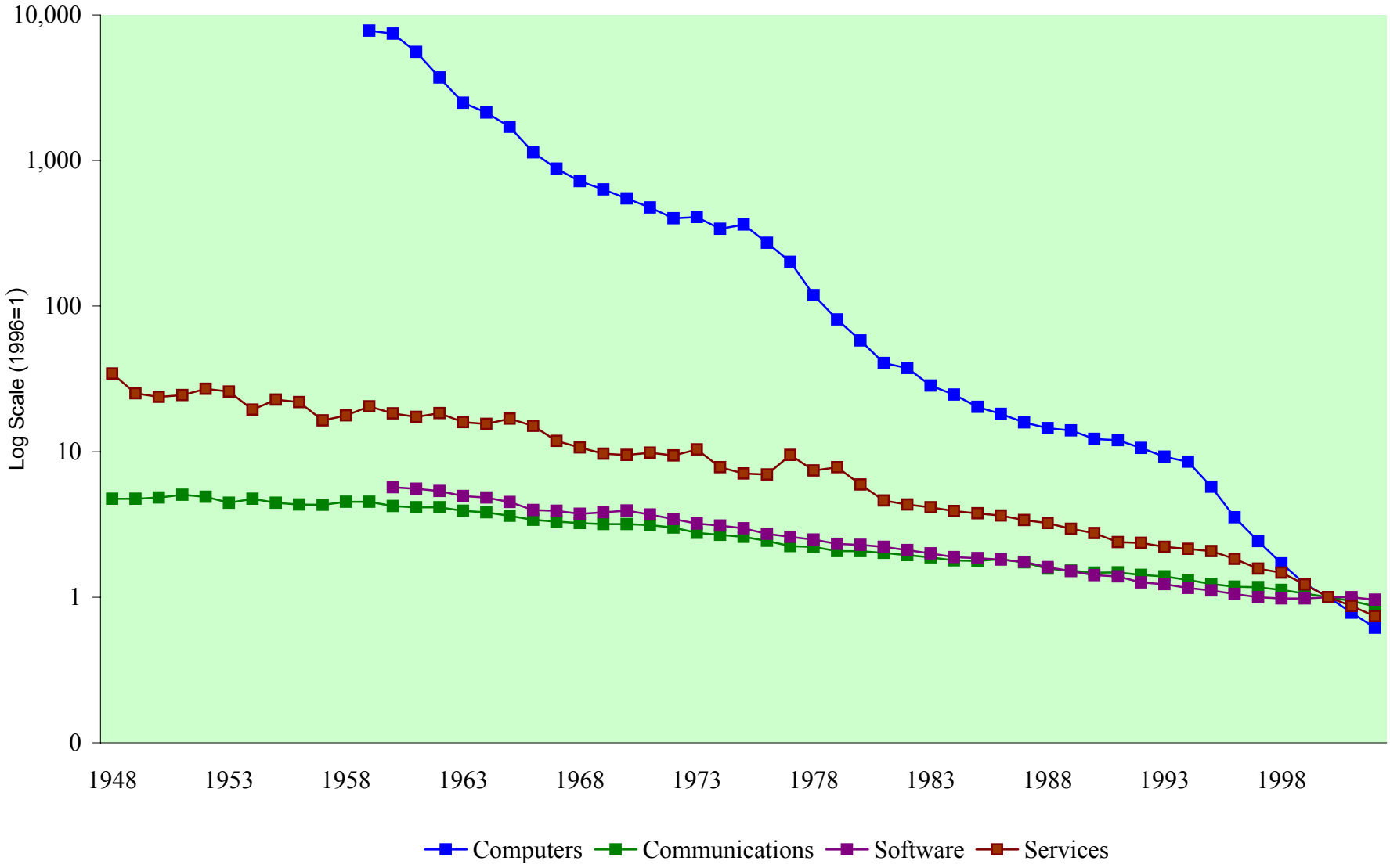
Note: Percentage. Contributions. Canada data begins in 1981

Figure 1.1: Relative Prices of Computers and Semiconductors, 1959-2002



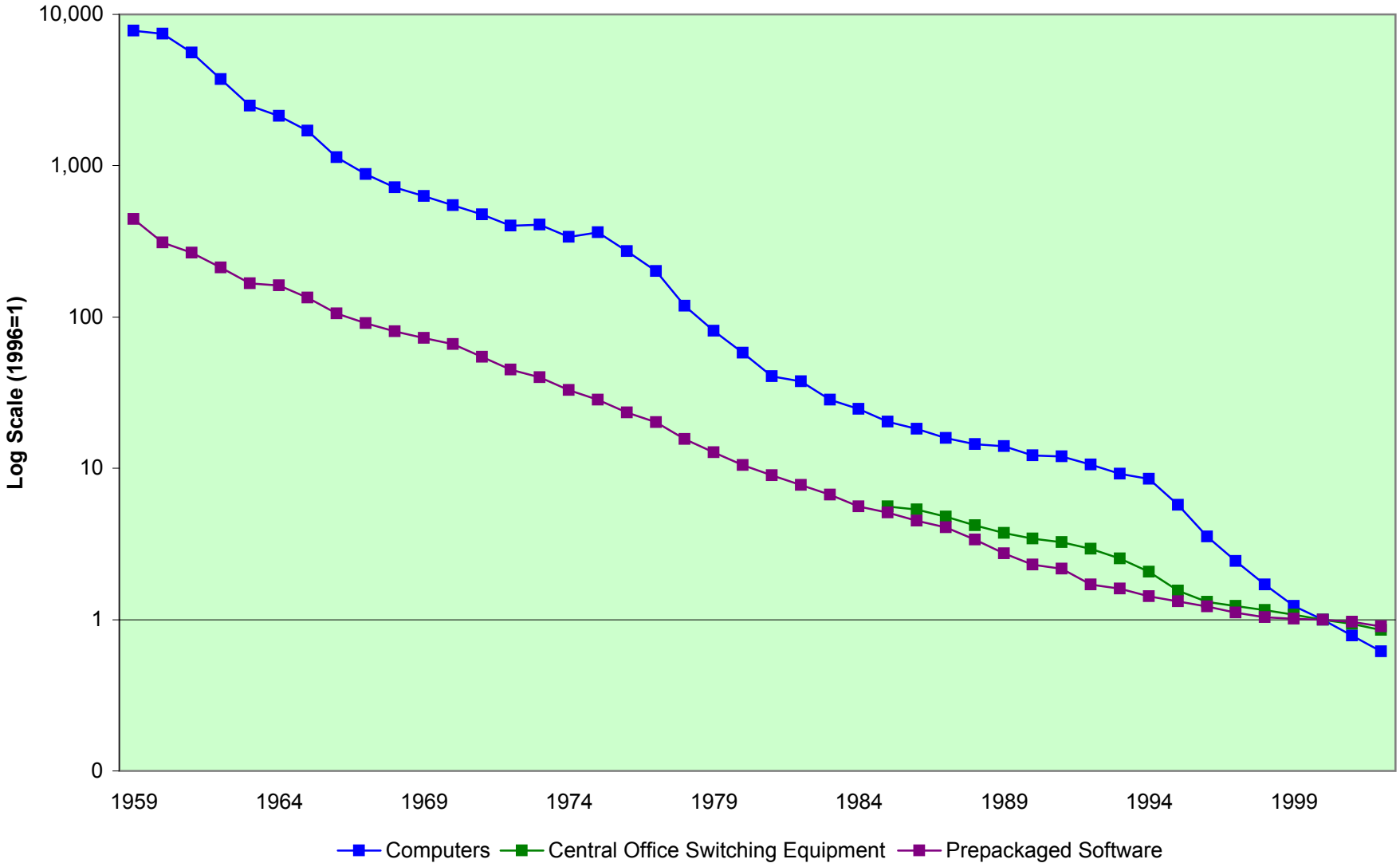
Note: All price indexes are divided by the output price index.

Figure 1.2: Relative Prices of Computers, Communications, Software, and Services, 1948-2002



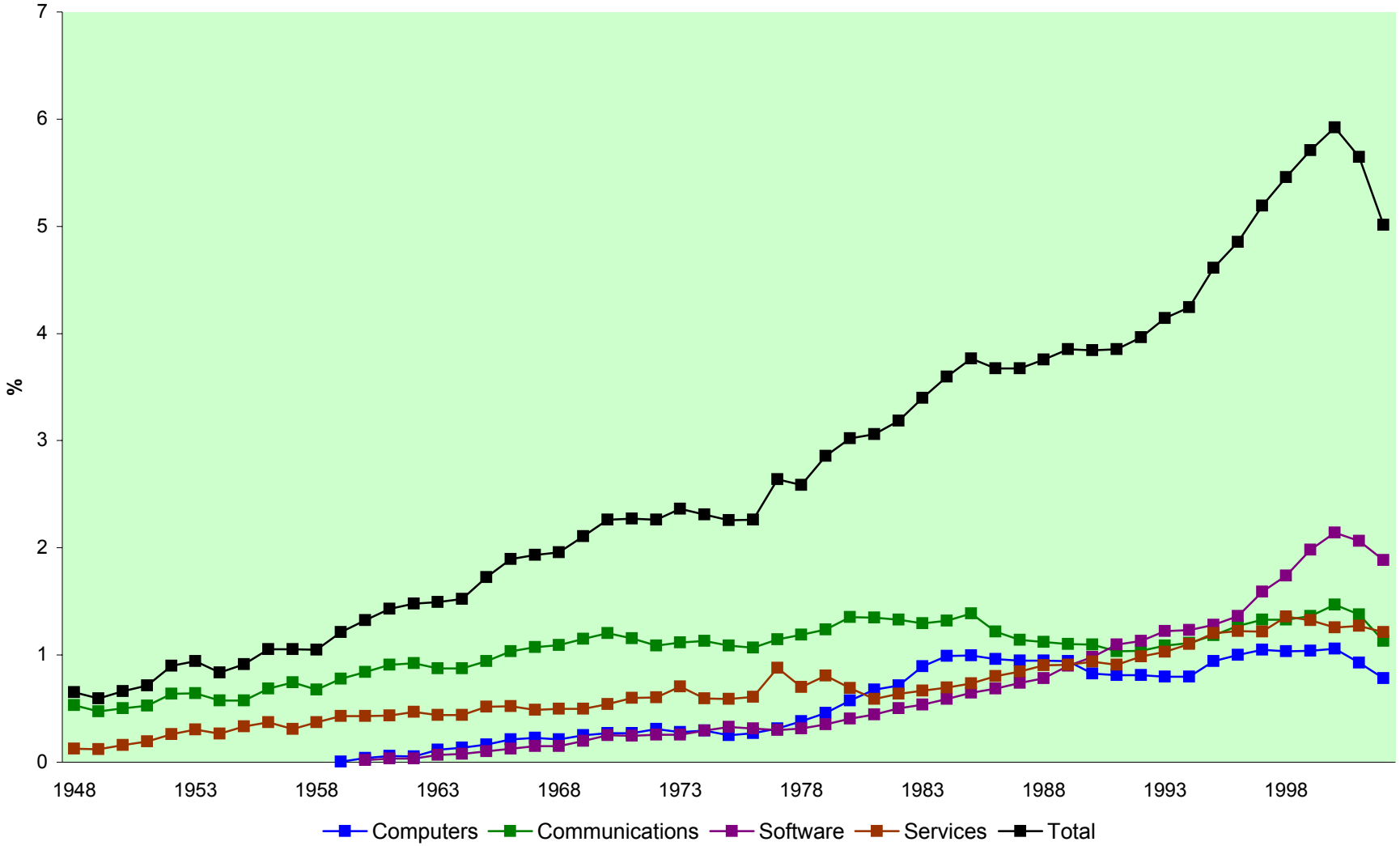
Note: All price indexes are divided by the output price index.

Figure 1.3: Relative Prices of Computers, Communications, and Software, 1959-2002



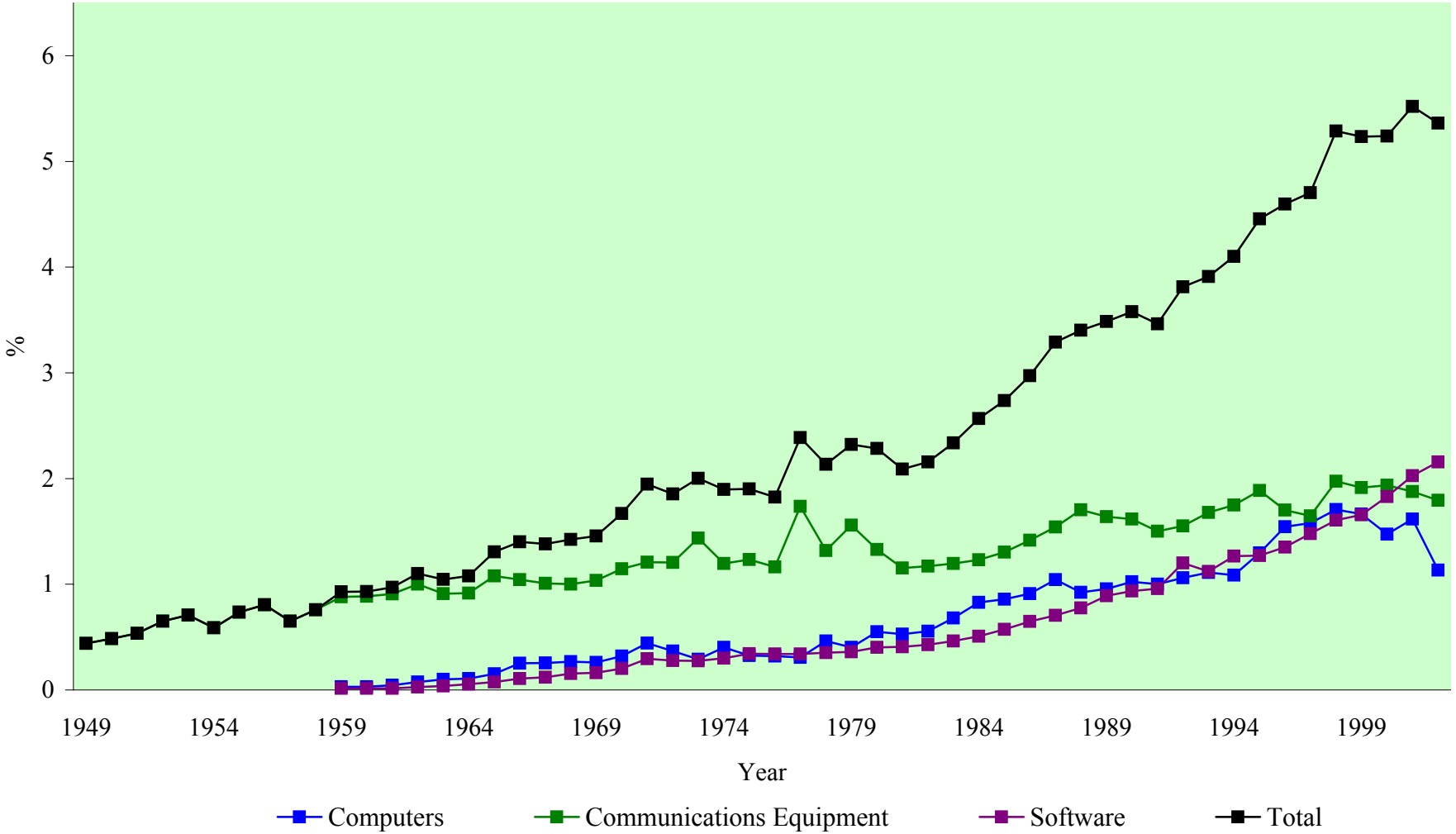
Note: All price indexes are divided by the output price index.

Figure 2.1: Output Shares of Information Technology by Type, 1948-2002



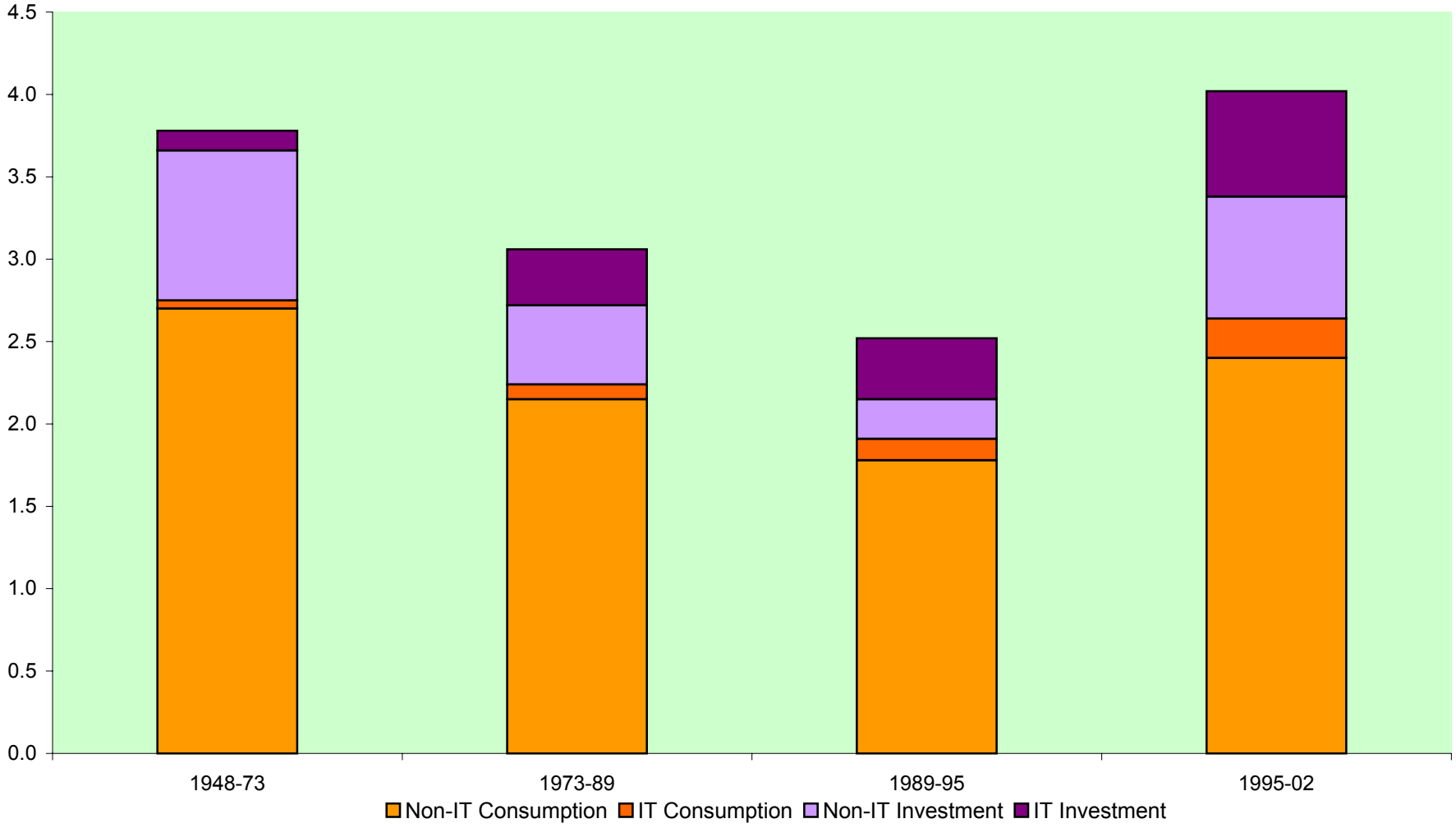
Note: Share of current dollar gross domestic product.

Figure 2.2: Input Shares of Information Technology by Type, 1948-2002



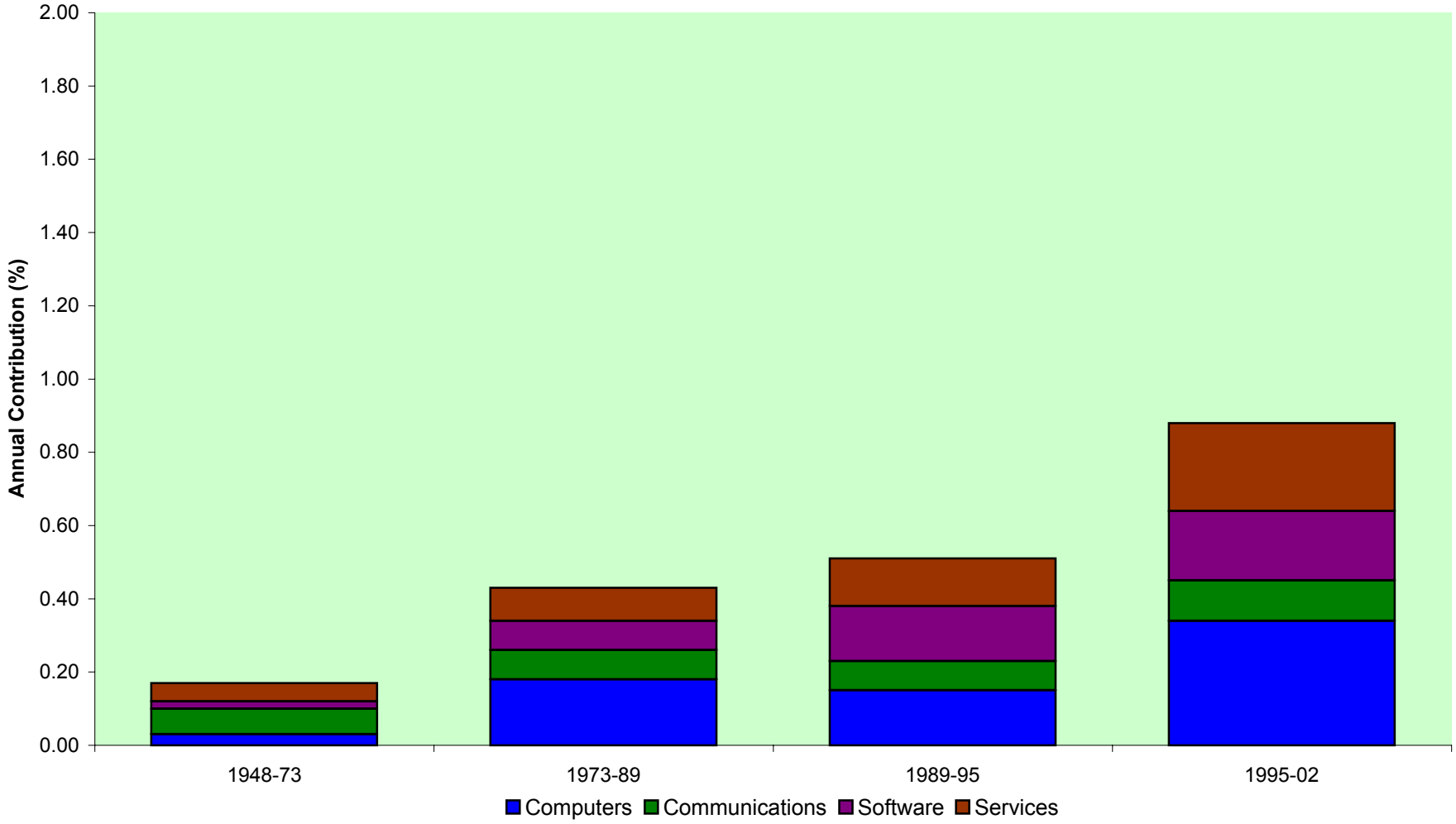
Note: Share of current dollar gross domestic income.

Figure 2.3: Output Contribution of Information Technology



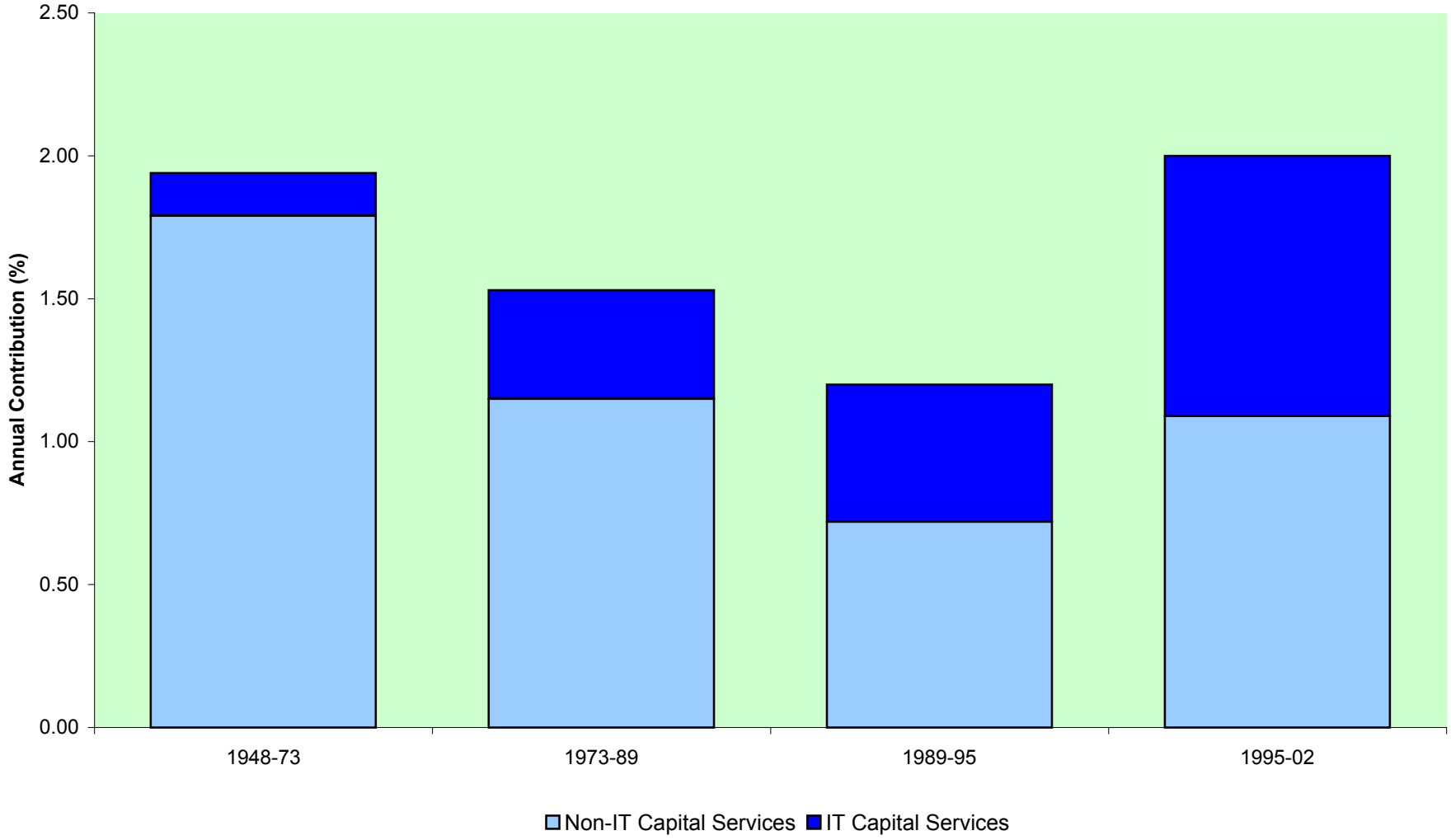
Note: Output contributions are the average annual growth rates, weighted by the output shares

Figure 2.4: Output Contribution of Information Technology by Type



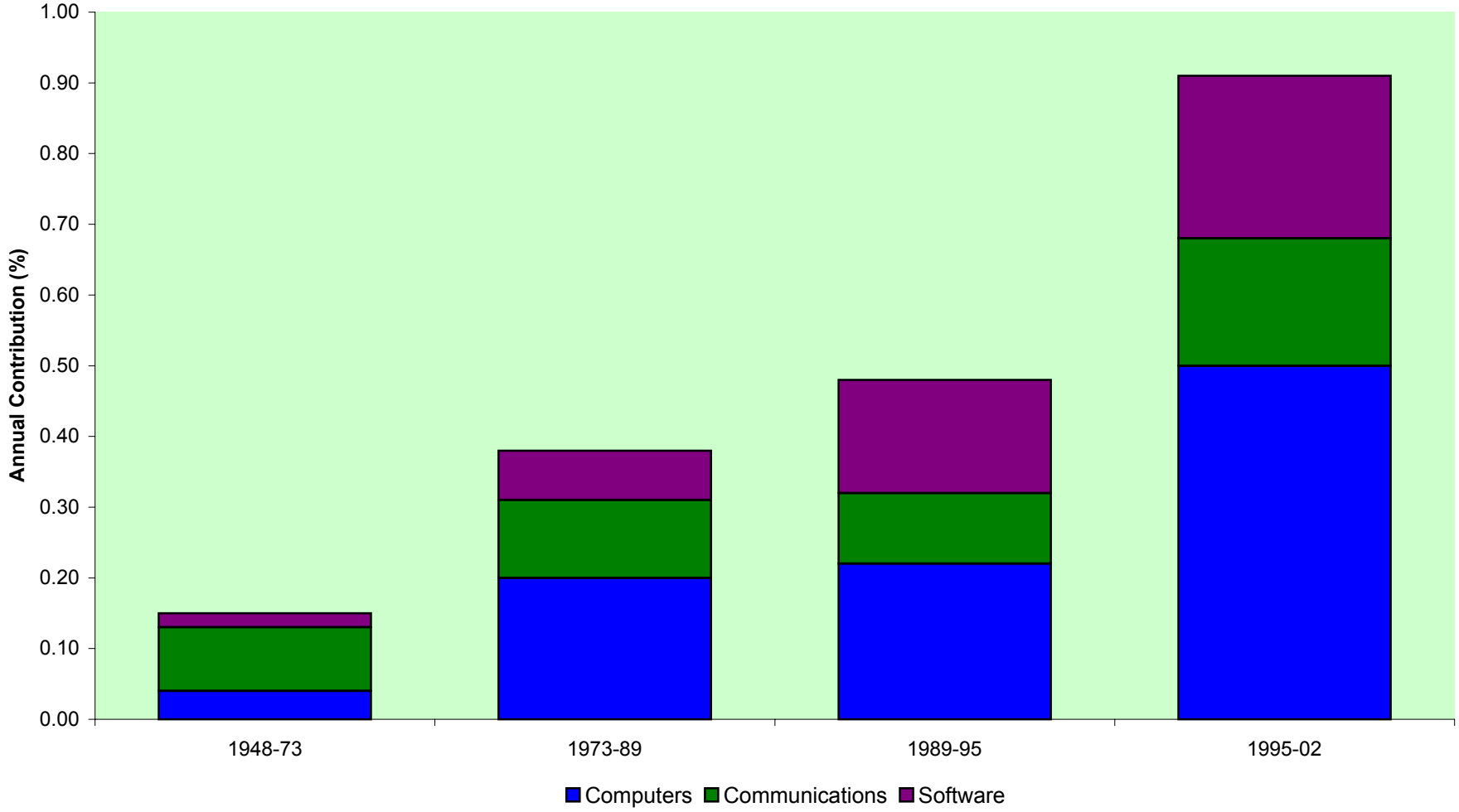
Note: Output contributions are the average annual growth rates, weighted by the output

Figure 2.5: Capital Input Contribution of Information Technology



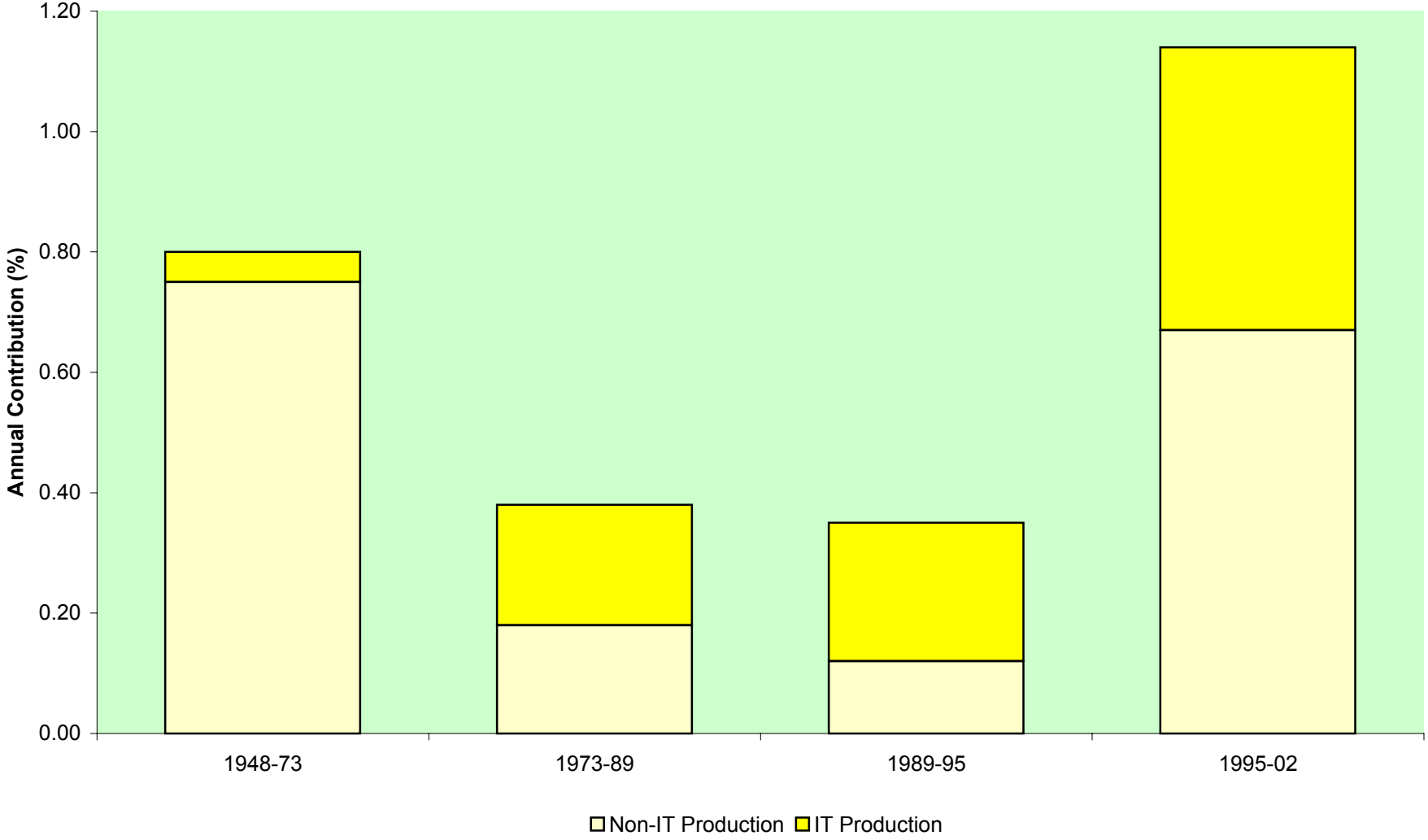
Note: Input contributions are the average annual growth rates, weighted by the income shares.

Figure 2.6: Capital Input Contribution of Information Technology by Type



Note: Input contributions are the average annual growth rates, weighted by the income shares.

Figure 2.7: Contributions of Information Technology to Total Factor Productivity Growth



Note: Contributions are average annual relative price changes, weighted by average nominal output shares from Table 8.

Figure 2.8: Sources of Gross Domestic Product Growth

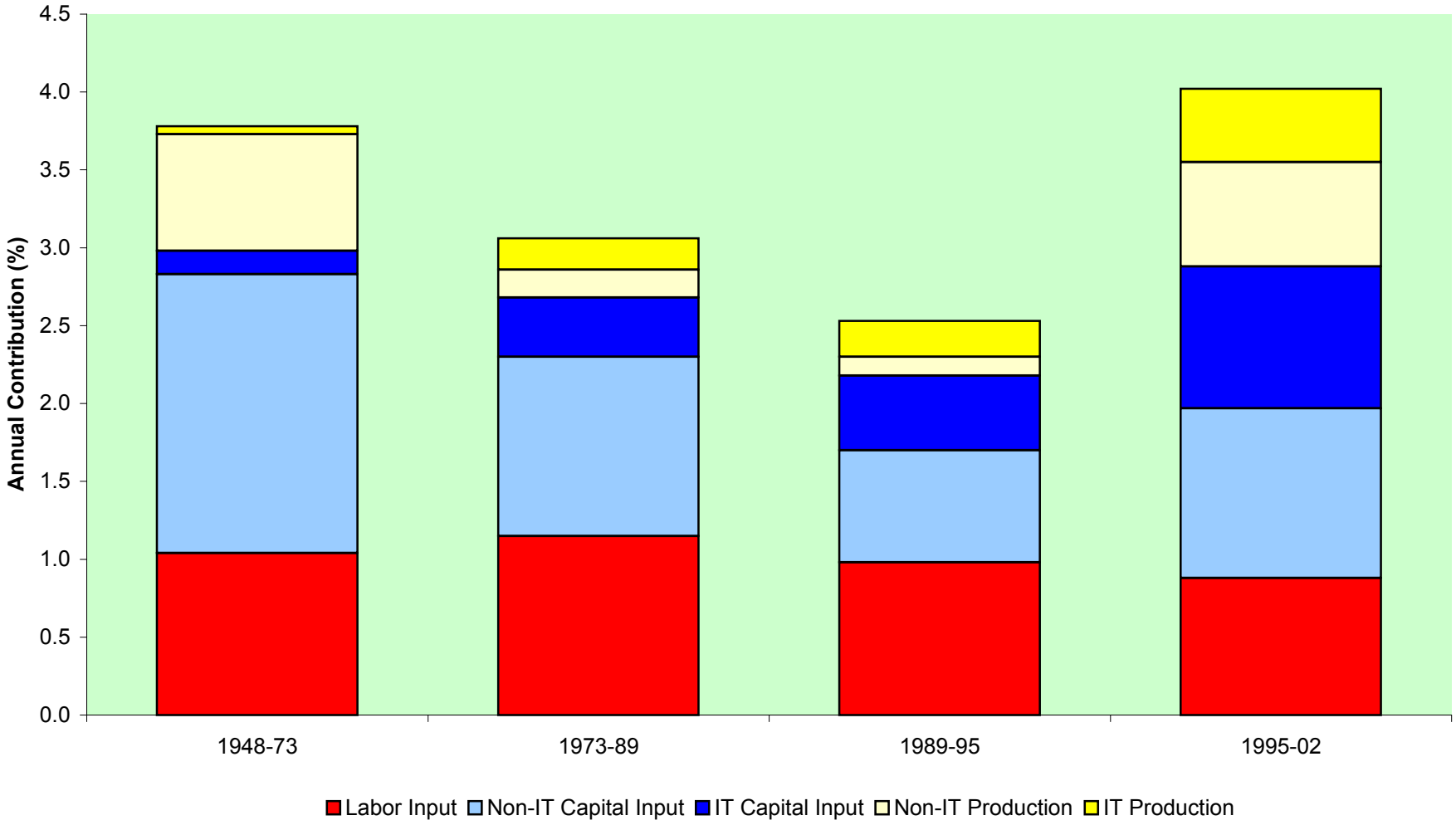


Figure 2.9: Sources of Average Labor Productivity Growth

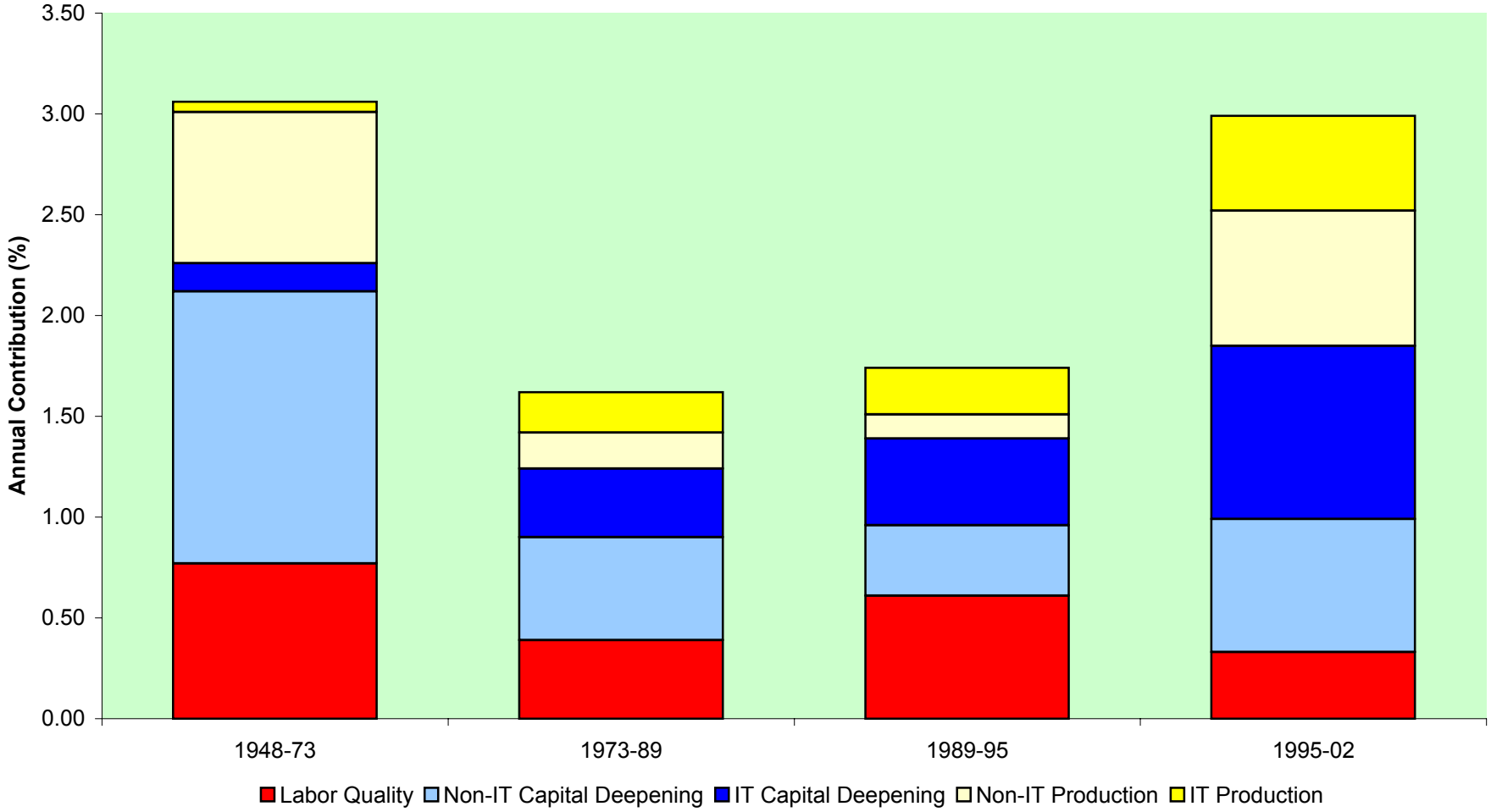


Figure 3.1 Capital Input Contribution by Country

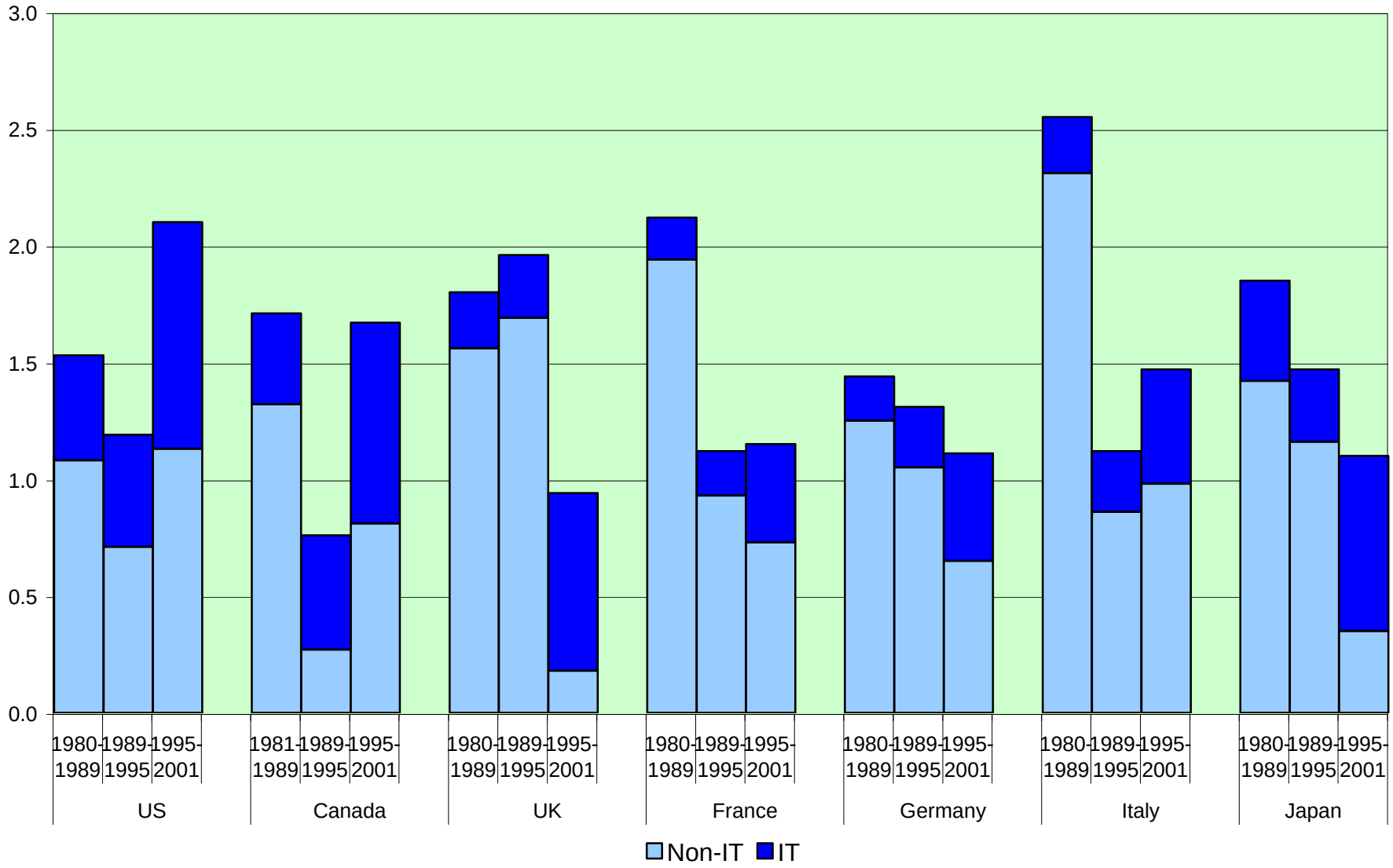


Figure 3.2 Sources of Productivity Growth by Country

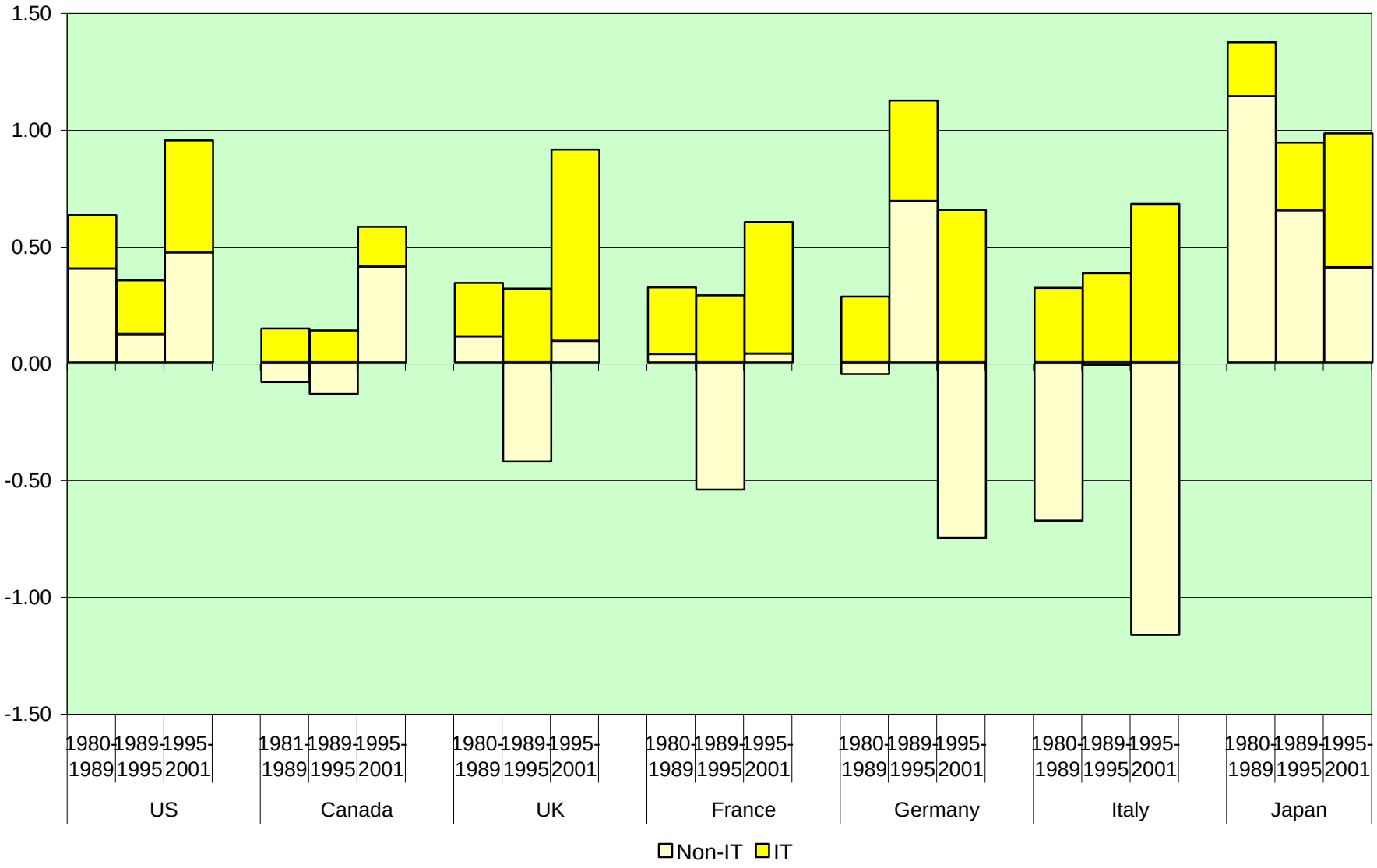


Figure 3.3 Sources of Economic Growth by Country

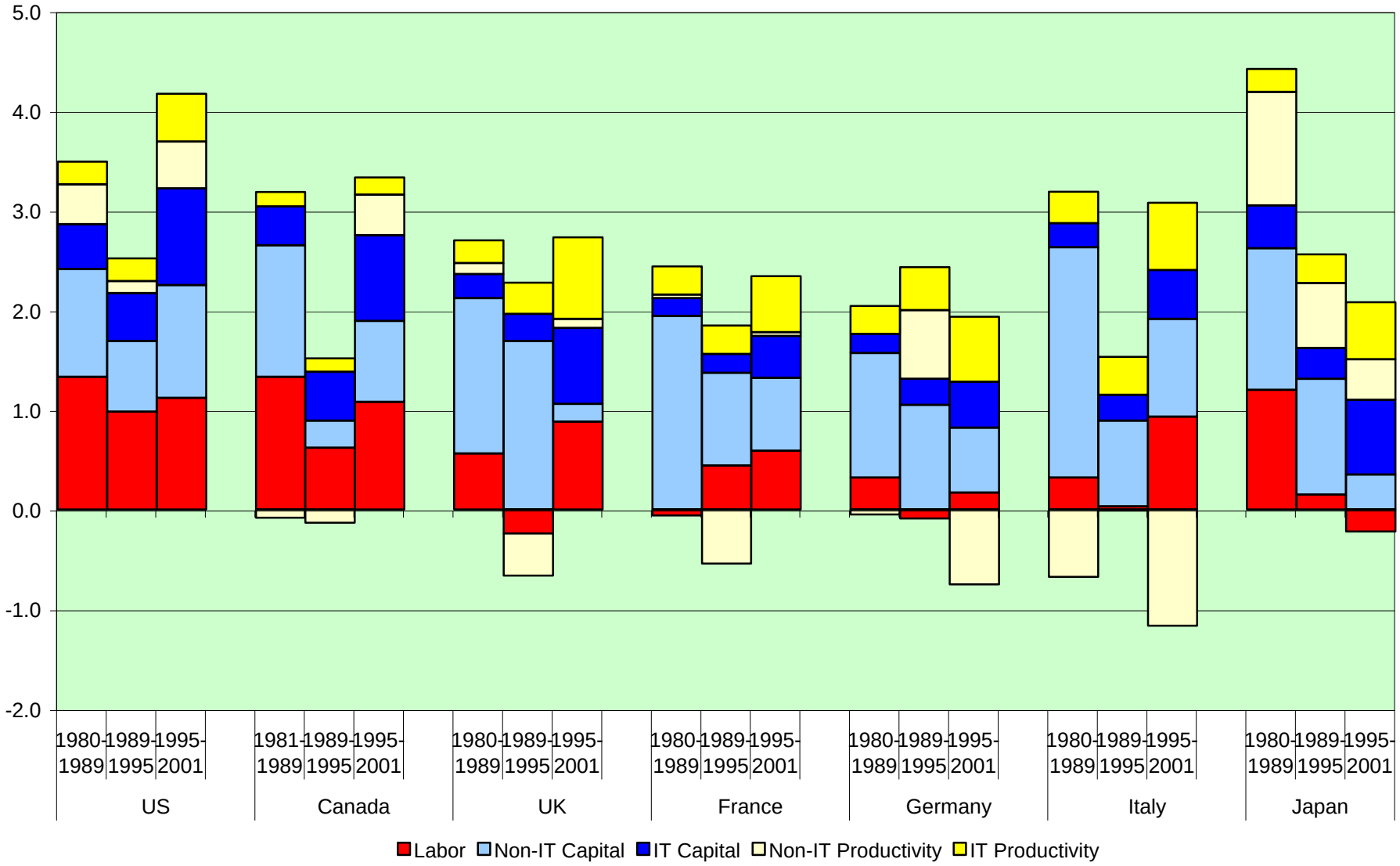
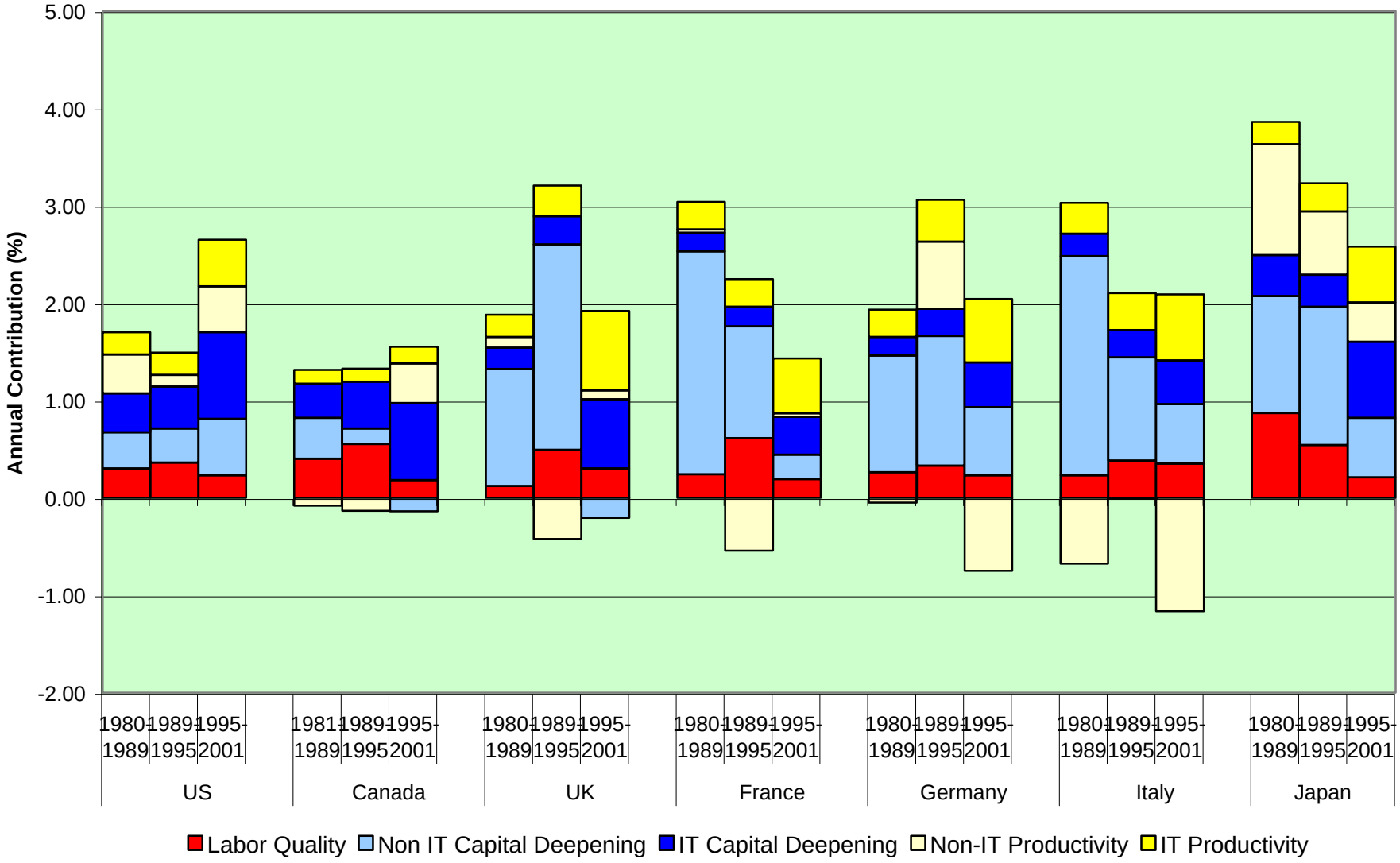


Figure 3.4 Sources of Labor Productivity Growth by Country



Externalities and Growth

Peter J. Klenow
Stanford University and NBER

Andrés Rodríguez-Clare
Inter-American Development Bank (IADB)

December 2004

Abstract

Externalities play a central role in most theories of economic growth. We argue that international externalities, in particular, are essential for explaining a number of empirical regularities about growth and development. Foremost among these is that many countries appear to share a common long run growth rate despite persistently different rates of investment in physical capital, human capital, and research. With this motivation, we construct a hybrid of some prominent growth models that have international knowledge externalities. When calibrated, the hybrid model does a surprisingly good job of generating realistic dispersion of income levels with modest barriers to technology adoption. Human capital and physical capital contribute to income differences both directly (as usual), and indirectly by boosting resources devoted to technology adoption. The model implies that most of income above subsistence is made possible by international diffusion of knowledge.

This is a preliminary and incomplete draft of a chapter prepared for the Handbook of Economic Growth, edited by Philippe Aghion and Steven Durlauf, North Holland Press, Amsterdam. For helpful comments we are grateful to Bob Hall, Chad Jones, Ben Malin, Paul Romer, and seminar participants at Berkeley, the IADB, MIT, Stanford and USC. Email: Pete@Klenow.com and AndresRo@IADB.org.

If ideas are the engine of growth and if an excess of social over private returns is an essential feature of the production of ideas, then we want to go out of our way to introduce external effects into growth theory, not try to do without them.

Robert E. Lucas (2002, p. 6)

1. Introduction

A number of facts suggest that international knowledge externalities are critical for understanding growth and development. The growth slowdown that began in the early 1970s was world-wide, not an OECD-only phenomenon. Countries with high investment rates exhibit higher income levels more than higher growth rates. Country growth rate differences are not very persistent from decade to decade, whereas differences in country incomes and investment rates are highly persistent. These patterns hold for investment rates in physical, human, and research capital. Together, they suggest that investment rates affect country transitional growth rates and long run relative incomes rather than long run growth rates. They also suggest countries are subject to the same long run growth rate. We argue that this represents evidence of very large international spillovers at the heart of the long run growth process.

We organize this chapter as follows. In Section 2 we describe two broad types of externalities and the growth models that do (and do not) feature them. Section 3 presents cross-country evidence that, we argue, is very hard to reconcile with the models that have no international externalities. Section 4 calibrates a model of growth with international externalities in the form of technology diffusion. The implied externalities are huge. Section 5 concludes and points out directions for future research.

2. A Brief Guide to Externalities in Growth Models

In this section we briefly discuss the role that externalities play in prominent theories of economic growth. One class of growth theories features externalities in the accumulation of knowledge possessed by firms (organizational capital) or by workers (human capital). Another class of growth models features externalities from the introduction of new goods, in the form of surplus to consumers and/or firms. Other theories

combine knowledge externalities and new good externalities. Finally, some important growth theories include no externalities at all. Table 1 provides examples of growth models categorized in these four ways. At the end of this section, we will dwell a little on the predictions of no-externalities models in order to motivate the evidence we describe in the next section. The evidence in the next section will suggest that models with no externalities cannot explain a number of empirical patterns.

2A. Models with Knowledge Externalities

Romer (1986) modeled endogenous growth due to knowledge externalities: a given firm is more productive the higher the average knowledge stock of other firms. As an example, consider a set of atomistic firms, each with knowledge capital *k,* benefiting from the average stock of knowledge capital in the economy *K* in their production of output *y*:

$$y_{it} = Ak_{it}^{\alpha} K_t^{1-\alpha}, \quad 0 < \alpha < 1. \tag{2.1}$$

Romer showed that, under certain conditions, constant returns to economy-wide knowledge, as in this example, can generate endogenous growth. The external effects are, of course, critical for long-run growth given the diminishing returns to private knowledge capital. Romer was agnostic as to whether the knowledge capital should be thought of as disembodied (knowledge in books) or embodied (physical capital and/or human capital).

Lucas (1988) was more specific, stressing the importance of human capital. Lucas sketched two models, one with human capital accumulated off-the-job and another with human capital accumulated on-the-job (i.e., learning by doing). Both models featured externalities. In the model with human capital accumulated off-the-job, Lucas posited

$$y_{it} = Ak_{it}^{\alpha} [u_{it} h_{it} n_{it}]^{1-\alpha} H_t^{\gamma}, \text{ with } \gamma > 0 \text{ and} \tag{2.2}$$

$$h_{it+1} = h_{it} + Bh_{it}[1 - u_{it}] \text{ with } 0 < u_{it} < 1. \tag{2.3}$$

Here u is the fraction of time spent working, and $1-u$ is the fraction of time spent accumulating human capital; h is an individual worker's human capital, and H is economy-

wide average human capital; *k* and *n* are physical capital and number of workers at a given firm. Because human capital accumulation is linear in the level of human capital, human capital is an engine of growth in this model. This is true with or without the externalities; *across*-dynasty externalities are not necessary for growth. As Lucas discusses, however, *within*-dynasty human capital spillovers are implicit if one imagines (2.3) as successive generations of finite-lived individuals within a dynasty. Within-dynasty externalities, however, would not have the same normative implications as across-dynasty externalities, namely underinvestment in human capital. Lucas (1988) did not argue that across-dynasty externalities were needed to fit particular facts. But he later observed that such across-household externalities could help explain why we see "immigration at maximal allowable rates and beyond from poor countries to wealthy ones" (Lucas 1990, p. 93).

Tamura (1991) analyzed a human capital externality in the production of human capital itself. This formulation conformed better to the intuition that individuals *learn* from the knowledge of others. Tamura specified

$$y_{it} = Ak_{it}^{\alpha}[u_{it}h_{it}n_{it}]^{1-\alpha} \tag{2.4}$$

$$h_{it+1} = h_{it} + B(h_{it}[1-u_{it}])^{\beta} H_t^{1-\beta}. \tag{2.5}$$

Because *H* represents economy-wide average human capital, $\beta < 1$ implies that learning externalities are essential for sustaining growth in Tamura's setup. If applied to each country, this model would suggest that immigrants from poor to rich countries should enjoy fast wage growth after they migrate, as they learn from being around higher average human capital in richer countries. Lucas (2004) used such learning externalities within cities as an ingredient of a model of urbanization and development.

Models not always thought of as having knowledge externalities are Mankiw, Romer and Weil's (1992) augmented Solow model and the original Solow (1956) neoclassical growth model. In Solow's model all firms within the economy enjoy the same level of TFP. This common level of TFP reflects technology accessible to all. The Solow model therefore does feature disembodied knowledge externalities across firms within an economy. In Mankiw et al.'s extension, knowledge externalities flow across countries as

well as across firms within countries. In section 4 we will discuss models with more limited international diffusion of knowledge. In these models imperfect diffusion means differences in TFP can play a role in explaining differences in income levels and growth rates. We stress that the Mankiw et al. model relies on even stronger externalities than the typical model of international technology spillovers, such as Parente and Prescott (1994) or Barro and Sala-i-Martin (1995, chapter 8). We will discuss these models at greater length in Section 4, when we calibrate a hybrid version of them.

2B. Models with Knowledge Externalities *and* New-Good Externalities

Models with both knowledge externalities and new-good externalities are the most plentiful in the endogenous growth literature. By "new-good externalities" we mean surplus to consumers and/or firms from the introduction of new goods. The new goods take the form of new varieties and/or higher quality versions of existing varieties. In Stokey (1988), learning by doing leads to the introduction of new goods over time. The new goods are of higher quality, and eventually displace older goods. The learning is completely external to firms, and what is learned applies to new goods even more than older goods. Hence learning externalities are at the heart of her growth process. In Stokey (1991), intergenerational human capital externalities (the young learn from the old) are critical for human capital accumulation. Human capital accumulation, in turn, facilitates the introduction of higher quality goods, which are intensive in human capital in her model.

Quality ladder models − pioneered by Grossman and Helpman (1991, chapter 4) and Aghion and Howitt (1992, 1998) − feature knowledge spillovers in that each quality innovation is built on the previous leading-edge technology. Such intertemporal knowledge spillovers are also fundamental in models with expanding product variety, such as Romer (1990) and Grossman and Helpman (1991, chapter 3). In Romer (1990),

$$Y = H_Y^{\alpha} L^{\beta} \int_0^A x(i)^{1-\alpha-\beta} di \tag{2.6}$$

$$\dot{A} = B H_A A. \tag{2.7}$$

Intermediate goods, the *x(i)*'s, are imperfect substitutes in production. This is the Dixit-Stiglitz "love of variety" model. The stock of varieties, or ideas, is *A*. In (2.7) new ideas are invented using human capital and, critically, the previous stock of ideas. This is the intertemporal knowledge spillover. Jones (1995, 2002) argues that, in contrast to (2.7), there are likely to be diminishing returns to the stock of ideas (an exponent less than 1 on *A*). He bases this on the fact that the number of research scientists and engineers have grown in the U.S. and other rich countries since 1950, yet the growth rate has not risen, as (2.7) would predict. Intertemporal knowledge spillovers still play a pivotal role in Jones' specification; they are just not as strong as in Romer's (2.7).

More recent models, such as Eaton and Kortum (1996) and Howitt (1999, 2000), continue to emphasize both knowledge externalities and new-good externalities. We will elaborate on these in Section 4 below.

2C. Models with New-Good Externalities

It is hard to find a model with new-good externalities but without knowledge externalities. We have identified three papers in the literature featuring such models, but two of the papers also have versions of their models with knowledge externalities.

Rivera-Batiz and Romer (1991) present a variation on Romer's (1990) model, as part of their analysis of the potential growth gains from international integration. In their twist, new intermediate goods are invented using factors in the same proportions as for final goods production in (2.6):

$$\dot{A} = BH^{\alpha}L^{\beta}\int_0^A x(i)^{1-\alpha-\beta}\,di\,. \tag{2.8}$$

They call this the "lab equipment model" to underscore the use of equipment in the research lab, just like in the production of final goods. In this formulation, they emphasize, "Access to the designs for all previous goods, and familiarity with the ideas and know-how that they represent, does not aid the creation of new designs" (p. 536-537). I.e., there are no knowledge externalities, domestic or international. Production of ideas is not even knowledge-intensive. Ideas are embodied in goods, however, and there is surplus to downstream consumers from their availability. Rivera-Batiz and Romer note that this

model allows countries to benefit from ideas developed elsewhere simply by importing the resulting products. Just as important, international trade allows international specialization in research. Countries can specialize in inventing different products, rather than every product being invented everywhere.

In a similar spirit, Romer (1994) considered a model in which knowledge about how to produce different varieties does not flow across countries, but each country can import the varieties that other countries know how to produce. For a small open economy, Romer posited

$$Y_t = A(\sum_{j=1}^{M_t} x_{jt}^{\alpha})N_t^{1-\alpha}, \quad 0 < \alpha < 1. \tag{2.9}$$

x_j represents the quantity of imports of the j^{th} variety of intermediate good. Because $\alpha < 1$, intermediate varieties are imperfectly substitutable in production. Firms in the importing country will have higher labor productivity the more import varieties they can access. If exporters cannot perfectly price discriminate and there is perfect competition among domestic final-goods producers, the higher labor productivity (higher *Y/N*) will benefit domestic workers/consumers. If consumer varieties were imported as well, there would be an additional source of consumer surplus from import varieties. Romer analyzed the impact of import tariffs on the number of varieties *M* imported in the presence of fixed costs of importing each variety in each country. Although Romer's model is static, growth in the number of varieties over time, say due to domestic population growth or falling barriers to trade, would be a source of growth in productivity and welfare in his model.

Kortum (1997) develops a model in which researchers draw techniques of varying efficiency levels from a Poisson distribution. Kortum does consider spillovers in the form of targeted search. But he also considers the case of blind search, wherein draws are independent of the previous draws. (Kortum fixes the set of goods produced but allows endogenous research into discovering better techniques for producing each good.) In the case of blind search, there are no knowledge spillovers. Growth is sustained solely because of population growth that raises the supply of and demand for researchers. It takes more and more draws to obtain a quality deep enough into the right tail to constitute an

improvement. A constant population growth rate sustains a constant flow of quality improvements and, hence, a constant growth rate of income.

2D. Models with No Externalities

The seminal growth models without externalities are the *AK* models of Jones and Manuelli (1990) and Rebelo (1991). In the next section we will present evidence at odds with such models, so we dwell on their implications here. We consider a version close to Rebelo's. Final output is a Cobb-Douglas function of physical and human capital:

$$C_t + I_t = Y_t = AK_{Yt}^{\alpha} H_{Yt}^{1-\alpha}, \tag{2.10}$$

where K_Y and H_Y represent the stocks of physical capital and human capital devoted to producing current output. As shown, current output can be used for either consumption or investment. The accumulation equations for physical and human capital are, respectively,

$$K_{Yt+1} + K_{Ht+1} = K_{t+1} = (1-\delta_K)K_t + I_t \tag{2.11}$$

$$H_{Yt+1} + H_{Ht+1} = H_{t+1} = (1-\delta_H)H_t + BH_{Ht}^{\gamma} K_{Ht}^{1-\gamma}. \tag{2.12}$$

H_H and K_H represent the stocks of human and physical capital, respectively, devoted to accumulating human capital.

We will focus on an equilibrium with a constant fraction of output invested in physical capital ($s_I = I/Y$) and a constant share of human capital deployed in human capital accumulation ($s_H = H_H / H$). We assume that the ratio of marginal products of physical and human capital are equated across the final output and human capital sectors, so that physical capital is devoted to

$$s_K = K_H / K = \frac{s_H(1-\gamma)(1-\alpha)}{s_H(1-\gamma)(1-\alpha) + (1-s_H)\gamma\alpha}. \tag{2.13}$$

The balanced growth rate is defined as

$$1+g = Y_{t+1}/Y_t = K_{t+1}/K_t = H_{t+1}/H_t. \tag{2.14}$$

The level of the balanced growth rate is an implicit function of the investment rates and parameter values:

$$(g+\delta_K)^{1-\gamma}(g+\delta_H)^{1-\alpha} = \left(A(1-s_K)^{\alpha}(1-s_H)^{1-\alpha}s_I\right)^{1-\gamma}\left(Bs_H^{\gamma}s_K^{1-\gamma}\right)^{1-\alpha}. \tag{2.15}$$

Provided $\alpha < 1$, human capital is the engine of growth. The growth rate is monotonically increasing in s_I because physical capital is an input into human capital accumulation whereas consumption is not. Related, the growth rate does not monotonically increase with the share of inputs devoted to producing human capital. This is because devoting more resources to producing current output increases the stock of physical capital, which is an input into human capital accumulation and hence growth.[1] When we look at the data in Section 3, however, we will find no country so high an s_H or s_K as to inhibit its growth according to this model.

When $\alpha = 1$ we have a literal $Y = AK$ model, and the growth rate is solely a function of the physical capital investment rate:

$$g+\delta_K = As_I \tag{2.16}$$

Here there is no point in devoting effort to producing human capital, so $s_H = 0$.

In the special case $\gamma = 1$, human capital is produced solely with human capital. This might be called a *BH* model. Presuming $\alpha < 1$ of course, the growth rate is simply

[1] To reinforce intuition, consider the (unrealistic) case of $\gamma = 0$, wherein new human capital is produced only with physical capital. In this case, growth is not strictly increasing in s_K (the share of capital devoted to human capital production) because some physical capital itself needs to be devoted to its own production.

(2.17) $$g + \delta_H = As_H$$

Unlike when $\gamma < 1$, the growth rate here is monotonically increasing in the effort devoted to adding more human capital. Lucas (1988) and many successors focus on this *BH* model because human capital accumulation is evidently intensive in human capital. Moreover, even *AK* models such as Jones and Manuelli (1990) construe their *K* to incorporate both human capital and physical capital. The consensus for diminishing returns to physical capital ($\alpha < 1$) is strong. Constant returns are entertained only for a broad measure of physical and human capital. We stress (2.15), a hybrid of *AK* and *BH* models, because this generalization allows us to take into account the combined impact of physical and human capital investment rates on growth when physical capital is an input to human capital accumulation ($\gamma < 1$).

3. Cross-Country Evidence

In this section we document a number of facts about country growth experiences over the last fifty years. We show that country growth rates appear to depend critically on the growth and income levels of *other* countries, rather than solely on domestic investment rates in physical and human capital. Cross-country externalities are a promising explanation for this interdependence. In brief, here are the main facts we will present:

- The growth slowdown that began in the mid-1970s was a *world-wide* phenomenon. It hit both rich countries and poor countries, and economies on every continent.

- Richer OECD countries grew much more slowly from 1950 to around 1980, despite the fact that richer OECD economies invested at higher rates in physical and human capital.

- Differences in country investment rates are far more persistent than differences in country growth rates.

- Countries with high investment rates tend to have high levels of income more than they tend to have high growth rates.

3A. The World-Wide Growth Slowdown

As has been widely documented for rich countries, the growth rate of productivity slowed beginning in the early 1970s.[2] Less widely known is that the slowdown has been a *world-wide* phenomenon, rather than just an OECD-specific event.[3] We document this in Table 2. Across 96 countries, the growth rate in PPP GDP per worker fell from 2.7% per year over 1960-1975 to 1.1% per year over 1975-2000. Growth decelerated 1.6 percentage points on average in both the sample of 23 OECD countries and the in the sample of 73 non-OECD countries.[4] The slowdown hit North and South America the hardest (their growth rates fell 2.4 percentage points) and barely brushed Asia (who slowed down just 0.4 of a percentage point). The slowdown hit all income quartiles of the 96 country sample (based on PPP income per worker in 1975). Although each income quartile grew at least one percentage point slower, the slowdown was not as severe in the poorest half as in the richest half. China's growth rate actually accelerated from 1.8 to 5.1, in the wake of reforms that began in the late 1970s. Chile, which experienced rapid growth in the 1990s, accelerated 2.1 percentage points.

Why does a world-wide growth slowdown suggest international externalities? Couldn't it simply reflect declining investment rates world-wide, as suggested by the *AK* model in the previous section? Table 2 also shows what average investment rates in physical and human capital did before and after the mid-1970s. The investment rates in physical capital come from Penn World Table 6.1. As a proxy for the fraction of time devoted to accumulating more human capital, we used years of schooling attainment relative to a 60-year working life. We used data on schooling attainment for the 25 and older population from Barro and Lee (2000). This human capital investment rate, which averages around 7% across countries, reflects the fraction of ages 5 to 65 devoted to schooling as opposed to working. We prefer the attainment of the workforce as opposed to

[2] The causes of the slowdown remain largely a mystery. For example, see Fischer (1988).
[3] An exception is Easterly (2001b).
[4] OECD countries are based on 1975 membership. There were 24 OECD members in 1975, but the Penn World Tables contain data for unified Germany only back to 1970.

the enrollment rates of the school-age population. The latter should take a long time to affect the workforce and therefore the growth rate.

According to Table 2, the average investment rate in physical capital across all countries was virtually unchanged (15.8% before vs. 15.5% after the slowdown), and the investment rate in human capital actually rose strongly (going from 7.1% to 9.7%). The same pattern applies for the OECD and non-OECD separately, and for all four quartiles of initial income. Thus the growth slowdown cannot be attributed to a world-wide decline in investment rates.

The breadth of the growth slowdown suggests *something* linking country growth rates, and ostensibly something other than investment rates.[5] This is contrary to the predictions of *AK* models, in which the growth rate of a country depends on domestic investment rates. The world-wide nature of the slowdown suggests that endogenous growth models, more generally, should not be applied to individual countries but rather to a collection of interdependent countries. Knowledge diffusion through trade, migration, and foreign direct investment are likely sources of interdependence.

Three other examples of interdependence are offered by Parente and Prescott (2004). First, growth rates picked up in the 20th century relative to the 19th century for many countries. Second, the time it takes a country to go from $2000 to $4000 in per capita income has fallen over the 20th century, suggesting the potential to grow rapidly by adopting technology in use elsewhere. Third and related, they stress that "growth miracles" always occur in countries with incomes well beneath the richest countries, again consistent with adoption of technology from abroad.

Knowledge diffusion, broadly construed, could include imitation of successful institutions and policies in other countries. Kremer, Onatski and Stock (2001) argue that such imitation might explain the empirical transition matrix of the world income distribution. If improving institutions leads only to static gains in efficiency, however, then the barriers to imitation have to be large to explain why the best institutions are not in place everywhere. As we will illustrate in section 4 below, the required barriers to technology adoption are modest precisely because the benefits accumulate with investments.

[5] It also casts doubt on explanations for the growth slowdown that are confined to rich countries.

3B. Beta Convergence in the OECD

As documented by Baumol (1986) and many others, incomes have generally been converging in the OECD. Barro and Sala-i-Martin (1995) used the term *sigma convergence* to describe such episodes of declining cross-sectional standard deviations in log incomes. We focus on a related concept that Barro and Sala-i-Martin labeled *beta convergence*, namely a negative correlation between a country's initial income level and its subsequent growth rate. We look at beta convergence year by year in Figure 1. The data on PPP income per worker comes from Penn World Table 6.1 (Heston, Summers and Aten, 2002), and covers 23 OECD countries over 1960-2000. The Figure shows the correlation between current income and growth hovering between –0.50 and –0.75 from 1960 through the early 1980s. The correlation was still negative from the mid-1980s through the mid-1990s, but less so, and turned positive in the latter 1990s.

De Long (1988) pointed out that a country's OECD membership is endogenous to its level of income, so that members at time *t* will tend to converge toward each other's incomes leading up to time *t*. Our focus, however, is not on convergence per se. Our point is instead about how investment rates correlate with income during the period of convergence. Figure 1 also shows the physical capital investment rate, and it is *positively* correlated with a country's income throughout the sample. Figure 2 shows that schooling attainment is also *positively* correlated with income throughout the sample.

How do these investment correlations square with simple *AK* models with no externalities? Expression (2.15) shows that a country's growth rate should be increasing in its investment rates. For beta convergence to occur in this model, a country's investment rates must be *negatively* correlated with a country's level of income. But Figures 1 and 2 show the opposite is true: in every year, richer OECD countries had *higher* investment rates in human and physical capital than poorer OECD countries did. According to this class of models, OECD countries should have been diverging throughout the entire sample, rather than converging through most of it. Now, this reasoning ignores likely differences in efficiency parameters *A* and *B* across countries. But rescuing *AK* models would require that richer countries have *lower* efficiency parameters. We would guess that rich countries tend to have better rather than worse institutions (e.g., Hall and Jones, 1999).

3C. Low Persistence of Growth Rate Differences

Easterly et al. (1993) documented that country growth rate differences do not persist much from decade to decade. They estimated correlations of around 0.1 to 0.3 across decades. In contrast, they found that country characteristics such as education levels and investment rates exhibit cross-decade correlations in the 0.6 to 0.9 range. Just as we do, they suggest country characteristics may determine relative income levels and world-wide technological changes long-run growth. Easterly and Levine (2001) similarly provide evidence that "growth is not persistent, but factor accumulation is."

In Table 3 we present similar findings. We compare average growth rates from 1980-2000 vs. 1960-1980, and from decade to decade within 1960-2000. We find growth rates much less persistent than investment rates for the world as a whole, and for the OECD and non-OECD separately. Again, these facts seem hard to reconcile with the *AK* model in which a country's domestic investment rates determine its growth rate.

Figure 3 illustrates a related pattern: deciles of countries (based on 1960 income per worker) grew at similar average rates from 1960 to 2000. Each decile consists of the unweighted average of income per worker in 9 or 10 countries. The average growth rate is 1.7% in the sample, and the bottom decile in 1960 grew at precisely this rate. This figure suggests movements in relative incomes, but no permanent differences in long-run growth rates, even comparing the richest and poorest countries. This sample contains 96 countries, and therefore many of the poorest countries mired in zero or negative growth.

Pritchett (1997), on the other hand, offers compelling evidence that incomes diverged massively from 1800 to 1960. Doesn't this divergence favor models, such as *AK* without international externalities, in which country growth rates are not intertwined? Not necessarily. As argued by Parente and Prescott (2004), the opening up of large income differences coincided with the onset of modern economic growth. The divergence could reflect the interaction of country-specific barriers to technology adoption with the emergence of modern technology-driven growth. More generally, any given divergence episode could reflect widening barriers to importing technology rather than simply differences in conventional investment rates.

3D. Investment Rates and Growth vs. Levels

The *AK* model we sketched in the previous section predicts that a country's growth rate will be strongly related to its investment rates in physical and human capital. In Table 4 we investigate this empirically in cross-sections of countries over 1960-2000. In four of the six cases, the average investment rate is positively and significantly related to the average growth rate. For the OECD, the physical capital investment rate is not significantly related to country growth, and the human capital investment rate is actually *negatively* and significantly related to country growth. But for the non-OECD and all-country samples, the signs and significance are as predicted. This evidence constitutes the empirical bulwark for *AK* models.

In the four cases where the signs are as predicted, are the magnitudes roughly as an *AK* model would predict? First consider the literal *AK* model. According to (2.16) in the previous section, the coefficient on s_I should be *A*. What might be a reasonable value for *A*? In order to match the average growth rate in GDP per worker (1.8%), given an average investment rate in physical capital (17%) and a customary depreciation rate (8%), the value of *A* would need to be

$$A = \frac{g^{avg} + \delta_K}{s_I^{avg}} = \frac{.018 + .08}{.17} \cong 0.57. \tag{3.1}$$

This level of *A* is more than four times larger than the two significant positive coefficients on s_I in the first column of Table 4, which are around 0.12. The estimated coefficients are small in magnitude compared to what an *AK* model would predict. This discrepancy could reflect classical measurement error in investment rates, but such measurement error would need to account for more than 80% of the variance of investment rates across countries. Plus one would expect *positive* endogeneity bias in estimating the average level of *A*, due to variation in *A* across countries that is positively correlated with variation in s_I.

We next consider the literal *BH* model. According to (2.17), the coefficient on s_H should be *B*. To produce the average growth rate in GDP per worker given the average

investment rate in human capital (8.8%) and a modest depreciation rate (2%), *B* would need to be

$$B = \frac{g^{avg} + \delta_H}{s_H^{avg}} = \frac{.018 + .02}{.088} \cong 0.43. \tag{3.2}$$

The third column of estimates in Table 4 contain coefficients on s_H. Of the two positive coefficients, one is half the predicted level (0.21) whereas the other is not far from the predicted level (0.37).

Finally, consider the hybrid model in (2.15). We assume $\gamma = 0.9$ so that human capital accumulation is intensive in human capital, but does use some physical capital. For producing current output we assume the standard physical capital share of $\alpha = 1/3$. We set the depreciation rates as previously mentioned. We set s_K, the share of physical capital devoted to human capital accumulation, based on (2.13). As (2.15) illustrates, we cannot independently identify *A* and *B*, only their product. We set $A^{1-\gamma}B^{1-\alpha} \cong 0.60,$ so that the average predicted growth rate from (2.15) and observed s_H and s_I investment rates matches the average growth rate in GDP per worker of 1.8%. We then regress actual growth rates on predicted growth rates for a cross-section of 73 countries with available data. The coefficient estimated is 0.26 (standard error 0.08, R^2 of 0.13), far below the theoretical value of 1. Again, the empirical estimate might be low because of measurement error in predicted growth, but it would need to be large.

To recap, only 1 of the 7 coefficients of growth on investment rates considered is in the ballpark of an *AK* model's prediction. In contrast, we obtain uniformly positive and significant coefficients when we regress (log) *levels* of country income on country investment rates. In 5 of the 6 cases, the R^2 is notably higher with levels than with growth rates. Investment rates appear far better at explaining relative income levels than relative growth rates. The driver of growth rates would appear to be something other than simply domestic investment rates.

The preceding discussion focused on the steady-state predictions of *AK* models. It is possible that *AK* models fare better empirically when transition dynamics are taken into

account. But it is worth noting that Klenow and Rodríguez-Clare (1997), Hall and Jones (1999), Bils and Klenow (2000), Easterly (2001a), Easterly and Levine (2001), and Hendricks (2002) all find that no more than half of the variation in growth rates or income levels can be attributed directly to human and physical capital. Pritchett (2004), who considers many different parameterizations of the human capital accumulation technology, likewise finds that human capital does not account for much cross-country variation in growth rates.

3E. R&D and TFP

We now turn away from *AK* models to a model with diminishing returns to physical and human capital, but with R&D as another form of investment. Such a model might be able to explain country growth rates with no reference to cross-country externalities. For example, perhaps a variant of the Romer (1990) model could be applied country by country, with no international knowledge flows. R&D investment would have to behave in a way that leads to a worldwide growth slowdown, beta convergence in the OECD, and low persistence of growth rate differences. And, more directly, R&D investment would have to explain country growth rates. Research effort, like human capital, is difficult to measure. But Lederman and Saenz (2003) have compiled data on R&D spending for many countries. We now ask the same questions of their R&D investment rates that we asked of investment rates in physical and human capital: how correlated are R&D investment rates with country growth rates and country income levels?

The first column in Table 5 says that countries with high R&D spending relative to GDP do not grow systematically faster.[6] Countries with high R&D shares do, however, tend to have high relative incomes. But the correlation with income is not significant outside the OECD. One possibility is that these regressions do not adequately control for the contributions of physical and capital. We therefore move to construct Total Factor Productivity (TFP) growth rates and levels. We subtract estimates of human and physical capital per worker from GDP per worker:

[6] Because R&D data was not available for all country-years between 1960 and 2000, we took time effects out of the variables (growth rates, income levels, investment rates in R&D), then averaged the residuals over time.

(3.3) $$\ln TFP = \ln(Y/L) - \alpha \ln(K/L) - (1-\alpha)\ln(H/L)$$

where Y is real GDP, L is employment, K is the real stock of physical capital, and H is the real stock of human capital. We suppress time and country subscripts in (3.3) for readability. We would prefer to let α vary across countries and across time based on factor shares, but such data is not readily available for most countries in the sample. We instead set $\alpha = 1/3$ for all countries and time periods. Gollin (2002) finds that capital's share varies from 0.20 to 0.35 across a sample of countries, but does not correlate with country income levels or growth rates. We use Penn World Table 6.1 data assembled by Heston, Summers and Aten (2002) for PPP GDP, employment, and PPP investment in physical capital. We assume an 8% geometric depreciation rate and the usual accumulation equation to cumulate investment into physical capital stocks. We approximate initial capital stocks using the procedure in Klenow and Rodríguez-Clare (1997, p. 78). We let human capital per worker be a simple Mincerian function of schooling:

(3.4) $$H = hL = \exp(\phi s)L .$$

Here h represents human capital per worker, and s denotes years of schooling attainment. We use Barro and Lee (2000) data on the schooling attainment of the 25 and older population. This data is available every five years from 1960 to 2000, with the last year an extrapolation based on enrollment rates and the slow-moving stock of workers. A more complete Mincerian formulation would include years of experience in addition to schooling and would sum the human capital stocks of workers with different education and experience levels. In Klenow and Rodríguez-Clare (1997) we found that taking experience and heterogeneity into account had little effect on aggregate levels and growth rates, so we do not pursue it here. We use (3.4) with the Mincerian return $\phi = 0.085$, based on the returns estimated for many countries and described by Psacharopoulos and Patrinos (2002).

The latter columns in Table 5 present regressions of TFP growth rates and levels on R&D investment rates. The sample of countries is smaller given data limitations (67 countries rather than 82). Just like growth in GDP per worker, growth in TFP is not significantly related to R&D investment rates. But TFP levels, like levels of GDP per

worker, are positively and significantly related to R&D investment rates. From this we conclude that even R&D investment rates affect relative income levels, not long-run growth rates. The persistence of R&D investment rate differences across countries, combined with the lack of persistent growth rate differences, supports this interpretation. We are led to consider models in which country growth rates are tethered together.

Before considering a model with international knowledge externalities, we pause to consider a model with "externalities" operating through the terms of trade. We have in mind Acemoglu and Ventura's (2002) model of the world income distribution. In their model, each country operates an *AK* technology, but uses it to produce distinct national varieties. Countries with high *AK* levels due to high investment rates plentifully supply their varieties, driving down their prices on the world market. This results in a *pAK* model with a stationary distribution of income even in the face of permanent differences in country investment rates (and *A* levels, for that matter). Prices tether incomes together in the world distribution, not the flow of ideas. This is a clever and coherent model, but we question its empirical relevance. Hummels and Klenow (2004) find that richer countries tend to export a given product at higher rather than lower prices. They do estimate modestly lower quality-adjusted prices for richer countries, but nowhere near the extent needed to offset *AK* forces and generate "only" a factor of 30 difference in incomes.

To summarize this section, *AK* models tightly connect investment rates and growth rates. Such a tight connection does not hold empirically. This is the case for the world growth slowdown, for OECD convergence, for growth persistence, and for country variation in growth vs. income levels. A version of the *AK* model with endogenous terms of trade might be able to circumvent these empirical hazards but faces empirical troubles of its own. We therefore turn to models with international knowledge externalities that drive long-run growth.

4. Models with common growth driven by international knowledge spillovers

Based on evidence in the previous section, we now focus on models with two features. The first is that, in steady state, all countries grow at the same rate thanks to international knowledge spillovers. The second feature is that differences in policies or

other country parameters generate differences in TFP levels rather than growth rates. Examples of this type of model are Howitt (2000), Parente and Prescott (1994), Eaton and Kortum (1996), as well as the model of technology diffusion in chapter 8 of Barro and Sala-i-Martin (1995).

In these models there is a world technology frontier, and a country's research efforts determine how close the country gets to that frontier. There are three different issues that must be addressed. First, what determines the growth rate of the world technology frontier? Second, how is it that a country's research efforts allow it to "tap into" the world technology frontier? And third, what explains differences across countries in their research efforts? Our goal in this section is to build on the ideas developed in the recent literature to construct a model that offers a unified treatment of these three issues and that is amenable to calibration. The calibration is intended to gauge the model's implications about the strength of the different externalities and the drivers of cross-country productivity differences.[7]

To highlight the different issues relevant for the model, our strategy is to present it in parts. The next subsection (4A) takes world growth and R&D investment as exogenous and discusses how R&D investment determines steady state relative productivity. Subsection 4B discusses different ways of modeling how world-wide R&D investment determines the growth rate of the world technology frontier. Subsection 4C extends the model so as to allow for endogenous determination of countries' R&D investment rates. Subsection 4D calibrates the model. Finally, subsection 4E presents the results of an exercise where we calculate, for each country in our sample, the impact on productivity from international spillovers.

4A. R&D investment and relative productivity

In this section we focus on a single country whose research efforts determine its productivity relative to the world technology frontier. Both the R&D investment rate and the rate of growth of the world technology frontier are exogenous. Output is produced with

[7] Although we refer to research externalities throughout this section, the knowledge externalities could just as well be with respect to human capital. Only when we use data on R&D is the analysis specific to research.

a Cobb-Douglas production function: $Y = K^{\alpha}(AhL)^{1-\alpha}$, where Y is total output, K is the physical capital stock, A is a technology index, h is human capital per person, and L is the total labor force. We assume that h is constant and exogenous. Output can be used for consumption (C), investment (I), or research (R), $Y = C + pI + R$, where p is the relative price of investment and is assumed constant through time. Capital is accumulated according to: $\dot{K} = I - \delta K$. Finally, A evolves according to:

$$\dot{A} = \left(\lambda R / L + \varepsilon A\right)\left(1 - A / A^*\right) \tag{4.1}$$

where λ is a positive parameter and A^* is the world technology frontier, both common across countries.[8]

There are three salient differences between this model and the standard endogenous growth model. Firstly, the productivity of research in generating A-growth is affected by the country's productivity relative to the frontier, as determined by the term $(1 - A/A^*)$ in (4.1). This captures the idea that there are "benefits to backwardness". One reason for this may be that the *effective cost* of innovation and technology adoption falls when a country is further away from the world technology frontier. This is what happens in Parente and Prescott (1994) and in Barro and Sala-i-Martin (1995, chapter 8). Alternatively, being further behind the frontier may confer an advantage because every successful technology adoption entails a *greater improvement* in the national technology level. This is what happens in Howitt (2000) and in Eaton and Kortum (1996).[9]

Secondly, we introduce $\varepsilon \geq 0$ to capture the sources of technology diffusion from abroad that do not depend on domestic research efforts. We have in mind imports of goods that embody technology, and that do not require upfront adoption costs (e.g, equipment

[8] In models like those of Parente and Prescott (1994) and Howitt (2000) research is meant to capture both R&D and technology adoption efforts. In this paper we follow this practice and simply refer to the sum of these two technology investments as R&D or just "research".

[9] In Howitt's model, $(1 - A/A^*)$ arises from the product of two terms: $(1/A^*)(A^* - A)$. The $(1/A^*)$ term arises because, as the world's technology becomes more advanced, more research is required to tap into it; the second term captures the fact that, when the country is more backward, every successful technology adoption entails a *greater improvement* in the national technology level.

which is no harder to use but which operates more efficiently).[10] As we will see below, this is important for the model to match certain features of the data.

Thirdly, in contrast to most endogenous growth models, we divide research effort by L in the *A*-growth expression above. This is done to get rid of scale effects and can be motivated in two ways. First, if *A* represents the quality of inputs, then one can envisage a process where an increase in the labor force leads to an expansion in the variety of inputs (Young, 1998 and Howitt, 1999). With a larger variety of inputs, research effort per variety is diluted. This eliminates the impact of *L* on *A* growth. Second, if research is undertaken by firms to increase their own productivity, then population growth may lead to an expansion in the number of firms and a decrease in the impact of aggregate research on firms' *A*-growth (Parente and Prescott, 1994). In this case, L represents the number of firms.

The measured R&D investment rate is given by $s_R = R/Y$. This implies that $R/(AL) = s_R Y/(AL) = s_R k$ where $k \equiv (K/Y)^{\alpha/(1-\alpha)} h = Y/(AL)$. To proceed, note that in steady state $a \equiv A/A^*$ will be constant, since *A* will grow at the same rate as *A**, which we denote by g_A. Thus, from (4.1)

$$g_A = \left(\lambda s_R k + \varepsilon\right)\left(1 - a\right) \tag{4.2}$$

Solving for a we obtain:

$$a = 1 - \frac{g_A}{\lambda s_R k + \varepsilon} \tag{4.3}$$

The values of k and s_R determine a country's relative A from (4.3). Conceivably, the parameter λ (TFP in research, if you will) could differ across countries and also contribute to differences in *A*. But in this paper we assume λ does not vary across countries. We do,

[10] This free flow of ideas is also likely to depend on the local presence of multinationals, which bring valuable knowledge that diffuses to other local firms without the need for additional R&D.

however, allow researchers to be more productive in countries with more physical and human capital per worker.

The previous results clearly show that policies that lower investment in physical or human capital or R&D do not affect a country's growth rate. Their effect is on a country's steady state relative *A*. Also, as discussed above, there are no scale effects in this model: higher *L* does not lead to higher growth or to a higher relative *A*. This stands in contrast to most growth models based on research (e.g., Romer 1990, Barro and Sala-i-Martin, 1995 – chapter 8).

It is also noteworthy in equation (4.3) that the value of k, which captures physical and human capital intensity, affects a country's TFP level conditional on its R&D investment rate. Thus, large differences in TFP across countries do not necessarily imply that differences in human and physical capital stocks are just a small part of cross country income differences. Indeed, this model suggests that some of the TFP differences may be due to differences in capital intensities across countries. Below we explore this issue quantitatively.

It is instructive to calculate the social rate of return to research at the national level. As shown in Jones and Williams (1998), this can be done even without knowing the details of the model that affect the endogenous determination of the R&D investment rate. Letting $\dot{A} = G(A, R)$, Jones and Williams show that the (within-country) social rate of return $\tilde{r}$ can be expressed as:

$$\tilde{r} = \frac{\partial Y / \partial A}{P_A} + \partial G / \partial A + g_{P_A}. \tag{4.4}$$

Here P_A stands for the price of ideas and is given by $P_A = \left(\partial G / \partial R\right)^{-1}$. As explained by Jones and Williams, the first two terms in (4.4) represent the dividends of research while the third term represents the associated capital gains. The first dividend term is the obvious component, namely the productivity gain from an additional idea divided by the price of ideas. The second dividend term captures how an additional idea affects the productivity of future R&D.

In the model we presented above, it is straightforward to show that, along a steady state path, we have:

$$\tilde{r} = (1-\alpha)\lambda k(1-a) + \left[\varepsilon(1-a) - \frac{a g_A}{1-a}\right] + g_L \tag{4.5}$$

The first term on the right-hand side corresponds to the first dividend term in Jones and Williams' formula. The second term, in square brackets, corresponds to the indirect effect of increasing *A* on the cost of research ($\partial G / \partial A$). The third term, g_L, corresponds to the term capturing the capital gains in Jones and Williams formula. To understand this last term, note that we have implicitly assumed that new varieties or firms start up with the same productivity as existing varieties or firms. Thus, the value of ideas will rise faster with a higher g_L, and the social return to research will correspondingly increase with g_L.

Also note that, since the RHS of (4.5) is decreasing in a and a is increasing in s_R, the social rate of return to research will be decreasing in s_R, as one would expect. If *k* varies less than a in the data, one should also expect to find higher social rates of return to research in poor countries than in rich countries, as found by Lederman and Maloney (2003).

More importantly, if ε is close to zero, then from (4.2) and (4.5) we should have $\tilde{r} \approx (1-\alpha)\lambda k(1-a) \approx (1-\alpha) g_A / s_R$. Using the growth rate of *A* in the OECD in the period 1960-2000 as an approximation of g_A (1.5%), and using $\alpha = 1/3$, then $\tilde{r} \approx 0.01 / s_R$. Noting that the median of s_R in the non-OECD countries we have in our sample is $s_R = 0.5\%$, then $\tilde{r} \geq 200\%$. This seems implausibly high.[11] There are two ways out of this problem. First, one can argue that measured R&D investment does not capture all the research efforts undertaken by countries. Clearly, higher R&D investment rates would lead to lower and more plausible social rates of return to research. Second, one can argue that the implausible implications of the model are due to the assumption that ε is close to zero. In the

[11] The problem is not so pronounced for the U.S. Given its measured R&D investment rate of $s_R = 2.5\%$, we have $\tilde{r} \approx 40\%$, which is in the range of estimates of the social rate of return to R&D in the U.S. See Griliches (1992) and Hall (1996).

calibration exercise in section 4D, we will argue that both of these solutions are needed to make the model consistent with the data.

4B. Modeling growth in the world technology frontier

In this section we extend the model so that g_A is endogenously determined by the research efforts in all countries. The models we mentioned above deal with this in different ways, except Parente and Prescott who leave g_A as exogenous. Barro and Sala-i-Martin (1995, chapter 8) have a Romer-type model of innovation that determines g_A in the "North." We do not pursue this possibility because of the scale effect that arises in their model (larger *L* in the North leads to higher g_A) and because we want to allow research efforts by all countries to contribute to the world growth rate. We first consider an adaptation of Howitt's (2000) formulation. A country's total effective research effort, λR_i, gets diluted by the country's number of varieties or number of firms, both represented by L_i, and is then multiplied by a common spillover parameter, σ, to determine that country's contribution to the growth of the world's technology frontier:

$$\dot{A}^* = \sigma \sum_i \left(\frac{\lambda R_i}{L_i} \right).$$

Given our results above, we obtain:

(H1) $$g_A = \sigma \sum_i \lambda k_i s_{Ri} a_i \, .$$

This formulation has the nice feature that the world growth rate does not depend on the world's level of *L* (no scale effect on growth at the world level), although it does depend positively on R&D investment rates. The main problem with this formulation, and the reason we do not pursue it further, is that larger countries contribute no more to world growth than smaller countries do. This has the implausible implication that subdividing countries would raise the world growth rate.

In footnote 21 of his paper, Howitt discusses an alternative specification wherein country spillovers are diluted by world variety rather than each country's variety. This implies that:

$$\dot{A}^* = \sigma \sum_i \left(\frac{\lambda R_i}{L} \right).$$

where $L = \sum L_i$. Howitt does not pursue this approach because, in the presence of steady-state differences in the rate of growth of L across countries, g_A would be completely determined in the limit by the research effort of the country with the largest rate of growth of L. We believe, however, that it is quite natural to analyze the case in which g_L is the same across countries.[12] In this case, $\omega_i \equiv L_i / L$ is constant through time, and the expression above can be manipulated to yield:

(H2) $$g_A = \sigma \sum_i \lambda k_i s_{Ri} a_i \omega_i \, .$$

If we think of L as the number of firms rather than the number of varieties of capital goods, then (H2) amounts to stating that g_A is determined by the country-workforce-weighted average research intensity across firms world-wide. This seems much more reasonable than (H1), where g_A is determined by the *unweighted* average of research intensity across countries.

Expression (H2) differs from (H1) only in the presence of the weights ω_i that represent shares of world L. This has two advantages: first, large countries contribute more to world growth than small countries do, and second, subdividing countries would not affect the world growth rate. But (H2) has a problematic implication, namely that those countries with higher than average $k_i s_{Ri} a_i$ would be better off disengaging from the rest of

[12] If one country's population did come to dominate world population, however, it might be sensible to say it does almost all of the world's research and, hence, it will virtually determine the world growth rate. We assume equal labor force growth rates across countries not because we think it is accurate for describing what is happening now, but because we think it is a convenient fiction for a steady state model to explore international spillovers.

the world – their growth rate would be higher if they were isolated. That is, a research intensive country would be better off ignoring the research done in other countries.

According to Howitt's variety interpretation of this model, this is because an isolated country's growth rate would be given by $\sigma\lambda k_i s_{Ri} a_i$. Its research intensity would no longer be spread out over the number of world varieties, but instead over the smaller number of the country's own varieties. Thus, when a country disengages, it no longer benefits from spillovers from research conducted by the rest of the world, but there is an important compensating gain that comes from the fact that variety – and therefore dilution – falls for the disengaging country. Since there is no love of variety in Howitt's model, a high research-intensity country would gain from disengagement. By this logic, engagement could not be sustained among any set of asymmetric countries! The higher $k_i s_{Ri} a_i$ countries would always prefer to disengage, leaving all countries isolated in equilibrium.

We now turn to an alternative specification for world spillovers in which variety does not play such a crucial role. The specification will exhibit several of the features we have been looking for: first, no scale effect of world population on the world's growth rate; second, other things equal, larger countries contribute more to world prosperity than small countries do; and third, tapping into rest-of-world research does not require spreading research across more varieties. We believe this is accomplished by adopting the formulation in Jones (1995): instead of dividing by L, the scale effect is avoided by introducing the assumption that advancing the world technology frontier gets harder as the frontier gets higher. This can be captured by the following specification of international spillovers:

$$\dot{A}^* = \left(A^*\right)^{\gamma-1} \sigma \sum_i \lambda R_i \tag{4.6}$$

where $\gamma < 1$. In this setting, sustained growth in A^* depends on a continuously rising population. To see this, notice that we can restate (4.6) as follows:

(J) $$g_A = \left(A^*\right)^{\gamma-1} L\sigma \sum_i \lambda k_i s_{Ri} a_i \omega_i \,.$$

This expression makes clear that g_A is decreasing in A^*; as mentioned above, this is what is going to eliminate the scale effect. Since all of the terms in the summation on the right-hand side of (J) are constant, then – differentiating with respect to time – we get that:

(4.7) $$g_A = \frac{g_L}{1-\gamma} \,.$$

One criticism of this specification is that g_A does not depend on s_R, hence policy-induced increases in research intensity would not increase the world's growth rate (Howitt, 1999). As Jones (2002) argues, however, research intensity has been increasing over the last decades without a concomitant increase in the growth rate, so it is far from clear that we want a model where g_A depends on s_R.[13]

An interesting and relevant feature of the model presented by Eaton and Kortum (1996) is that it allows for spillovers to differ between pairs of countries. We can introduce this feature in the model by doing two things: first, we allow each country to have a different technology frontier, A_i^*; second, we add country-pair specific spillover parameters, η_{il}, to (4.6) so that now:

$$\dot{A}_i^* = \left(A_i^*\right)^{\gamma-1} \sigma \sum_l \lambda R_l \eta_{il} \,.$$

This formulation implies that there will no longer be a world technology frontier in the way it existed in model (J). However, it proves useful for the analysis to introduce a new concept, which we will denote by $\tilde{A}$ and which could be understood as the "frictionless technology frontier." To define this concept, note that if spillovers were the same among

[13] Even though research intensity does not affect the growth rate, it can have sizable effects on welfare, particularly when – as evidence suggests – the social rate of return of research is significantly higher than the private rate of return.

all country pairs ($\eta_{il} = 1$ for all *i* and *l*) – a case we could interpret as frictionless – then countries would have a *common* technology frontier: $A_i^* = A_l^*$ for all *i* and *l*. We define $\tilde{A}$ so that in this case ($\eta_{il} = 1$ for all *i* and *l*) $A_i^* = \tilde{A}$ for all *i*. As we will see below, in steady state $\tilde{A}$ grows at the same rate as A_i^* for all *i*. Letting $z_i \equiv A_i^* / \tilde{A}$, which captures the strength of spillovers from the rest of the world to country *i*, we arrive at the following steady state restriction:

(JEK) $$g_A = \left(A_i^*\right)^{\gamma-1} L(\sigma / z_i) \sum_l \lambda k_l s_{Rl} a_l \omega_l z_l \eta_{il}$$

where JEK stands for Jones, Eaton and Kortum and where a_l is now country *l*'s technology level relative to *its own* technology frontier: $a_l \equiv A_l / A_l^*$. It can be shown that this implies the following restriction for $\tilde{A}$:

(4.8) $$\tilde{A} = (vL)^{1/(1-\gamma)}$$

where $v \equiv (\sigma / g_A) \sum_l \lambda k_l s_{Rl} a_l \omega_l$. It is clear that each country's technology frontier and $\tilde{A}$ will grow at the same rate as A^* did in model (J), given by $g_L /(1-\gamma)$.

It is interesting to pause here to discuss the model's implications regarding the effect of country size on productivity. Imagine, to simplify, that all countries are the same except for size, and assume that $\eta_{ij} = 1$ for $i = j$ and $\eta_{ij} = \eta < 1$ for $i \neq j$. Then it is easy to show that $z_i > z_j$ if $\omega_i > \omega_j$; larger countries are more productive. Intuitively, larger countries benefit more from spillovers because more of the world's research takes place within their borders. As long as borders discretely reduce spillovers, larger countries will capture more spillovers and enjoy higher productivity.

The next step is to impose some restrictions on the international spillover parameters η_{il}'s. The literature has allowed international spillovers to depend on trade (Coe and Helpman, 1995), distance (Eaton and Kortum, 1996), and other variables such as FDI

flows (Caves, 1996). Here we focus on the simplest approach, which is to assume that the parameters η_{il} are completely determined by distance. (This would capture trade and FDI related spillovers that are related to distance.) We do this by assuming that $\eta_{il} = e^{-\theta d(i,l)}$, where *d(i,l)* is bilateral distance between countries *i* and *l*, and θ is some positive parameter. This model collapses to (J) if $\theta = 0$.

This completes our discussion of different ways to model international spillovers. Table 6 summarizes the discussion in this subsection.

4C. Determinants of R&D investment

We mentioned above that there are two ways to motivate the model we presented in subsection 4A. First, we can think of a model like the one presented in Howitt (2000), where research leads to improvements in the quality of capital goods, and population growth leads to an expansion in the total number of varieties available. Second, research may be carried out by firms to increase their own productivity, as in Parente and Prescott (1994). We pursue this second approach because it is simpler and much more convenient for our calibration purposes later on.

As in Parente and Prescott (1994), we assume a constraint on the amount of labor firms can hire. In particular, we assume that firms can hire no more than *F* workers. To simplify notation, we set *F* = 1. This constraint can be motivated as a limitation on the span of control by managers, as in Lucas (1978).[14] Output produced by firm *j* in country *i* at time *t*, which we denote by Y_{jit}, is given by $Y_{jit} = K_{jit}^{\alpha}(A_{jit}h_i)^{1-\alpha}$. (We now use time subscripts because they clarify the maximization problem below.) The firm can convert output into consumption, investment goods or R&D according to $Y_{jit} = C_{jit} + p_i I_{jit} + R_{jit}$, and the firm's capital stock evolves according to $\dot{K}_{jit} = I_{jit} - \delta K_{jit}$. Finally, the firm's technology index A_{jit} evolves according to:

$$\dot{A}_{jit} = \left((1-\mu)\lambda R_{jit} + \mu\lambda \bar{R}_{it} + \varepsilon A_{jit}\right)\left(1 - A_{jit}/A_{it}^{*}\right) \tag{4.9}$$

[14] If one takes *F* = 1 literally, then the externalities are in the human capital investment of individual workers.

where μ is a parameter between zero and one, $\bar{R}_{it}$ is the average of R_{jit} across firms in country *i* (we use the bar over the variable to emphasize that this is the average across firms, and not the aggregate economy-wide variable), and A^*_{it} is the technology frontier for country *i* with $\dot{A}^*_{it} / A^*_{it} = g_A$ for all *i in steady state.*

There are two features in this specification that merit some explanation. First, the "benefits of backwardness" are determined by the term $1 - A_{jit} / A^*_{it}$, which can differ across firms in country *i*: a more backward firm in country *i* would have a higher catch-up term. If instead we specified the catch-up term as $1 - \bar{A}_{it} / A^*_{it}$ (where $\bar{A}_{it}$ is the average technology index across firms in country *i*), then there would be a negative externality because, as a firm does more research, it increases the country's average technology index and decreases the catch-up term for the other firms. Given that there is no particular reason to think that this negative externality is a relevant feature to include in the model, we have chosen to specify the catch-up term as $1 - A_{jit} / A^*_{it}$. Second, there is a positive research externality across firms within each country, represented by the term $\mu\lambda\bar{R}_{it}$. This externality captures the idea that a firm benefits directly from research undertaken by other firms within the same economy.

To relate this to what we had in subsection 4A, note that if firms within a country are identical, then $R_{jit} = \bar{R}_{it}$ and $A_{jit} = \bar{A}_{it}$. Using this in (4.9), we obtain:

$$\dot{\bar{A}}_{it} = \left(\lambda \bar{R}_{it} + \varepsilon \bar{A}_{it}\right)\left(1 - \bar{A}_{it} / A^*_{it}\right)$$

But note that $A_{it} = \bar{A}_{it}$ and $\bar{R}_{it} = R_{it} / L_{it}$, where L_{it} is the total labor force in country *i* and also the number of firms there, given our assumptions above. Using these results and noting that $s_{Ri} = R_{it} / Y_{it}$ we obtain equation (4.2).

Firms in country *i* pay taxes at rate τ_{Ki} on capital income (output less the wage bill) ,and there is an R&D tax (or subsidy, if it is negative) of τ_{Ri}.[15] We stress that this R&D tax parameter does not have to be interpreted strictly as a formal tax or subsidy; when positive, the R&D tax parameter τ_{Ri} could also be interpreted as capturing "barriers to technology adoption", as in Parente and Prescott (1994).[16]

The firm's dynamic optimization problem is to choose a path for R_{jis} and I_{jis} to maximize

$$\int_t^{\infty} \left((1-\tau_{Ki}) \left[Y_{jis} - w_{is} \right] - p_i I_{jis} - (1+\tau_{Ri}) R_{jis} \right) e^{-r(s-t)} ds$$

subject to $\dot{K}_{jis} = I_{jis} - \delta K_{jis}$, $\dot{A}_{is} / A_{is} = \dot{A}^*_{is} / A^*_{is} = g_A$, and

$$\dot{A}_{jis} = \left((1-\mu)\lambda R_{jis} + \mu\lambda \bar{R}_{is} + \varepsilon A_{jis} \right) \left(1 - A_{jis} / A^*_{is} \right)$$

As shown in the Appendix, by imposing the symmetry condition on the two Euler equations for this optimization problem, we obtain the following two conditions for the symmetric equilibrium:

(4.10) $$\frac{p_i K_{it}}{Y_{it}} = \alpha \frac{1-\tau_{Ki}}{r+\delta}$$

(4.11) $$\Omega_i (1-\alpha)\lambda k_i (1-a_i) - g_A a_i / (1-a_i) + \varepsilon (1-a_i) = r$$

where $$\Omega_i \equiv \frac{(1-\tau_{Ki})(1-\mu)}{(1+\tau_{Ri})}.$$

[15] We should note here that the tax rate on capital income also affects the incentive to do research. The notation used for the two tax rates is meant to emphasize that τ_{Ki} affects all forms of accumulation by the firm, whereas τ_{Ri} only affects research expenditures.

[16] We assume that any tax revenue collected is distributed back to consumers in lump-sum fashion.

Equation (4.10) defines the equilibrium capital-output ratio in country *i* and equation (4.11) implicitly defines the equilibrium relative A in country *i*. Given a_i and knowing k_i from the data, we can plug their values into equation (4.3) to obtain the equilibrium steady state R&D investment rate, s_{Ri}. It is easy to see that an increase in the capital income tax or the R&D tax or an increase in the externality parameter, μ, would decrease Ω_i and hence lead to a decline in equilibrium a_i (this is because the left-hand side of (4.11) is decreasing in a_i). This, of course, would imply a decline in the R&D investment rate. The same reasoning shows that a_i is increasing in k_i but it is not necessarily the case that s_{Ri} increases with k_i (see the Appendix).

Combining the result for the social rate of return in equation (4.5) with (4.11), we obtain the following expression for the wedge between the social and private rate of return to R&D:

$$\tilde{r}_i - r = (1-\Omega_i)(1-\alpha)\lambda k_i(1-a_i) + g_L \tag{4.12}$$

The first term on the right-hand side is the distortion created by Ω, which captures the effect of the income tax, τ_K, the R&D tax, τ_R, and the externality parameter, μ. If there are no taxes and $\mu = 0$ (no domestic R&D externalities), then $\Omega_i = 1$ and the wedge between the social and private rate of return to R&D collapses to g_L.[17]

4D. Calibration

The model described in the previous section, together with the (JEK) formulation for international spillovers with $\eta_{il} = e^{-\theta d(i,l)}$, constitutes the model we calibrate in this subsection. Since we will only be working with the symmetric steady state equilibrium, in this subsection we suppress time and firm subscripts to simplify notation. Given N

[17] As explained above, g_L is associated with a positive externality because new firms start up with the same productivity as existing firms. Since the number of firms is equal to the workforce, the value of ideas and the social rate of return are increasing in g_L.

countries, the steady state equilibrium is given by $\{K_i, Y_i, k_i, a_i, s_{Ri}, A_i, A_i^*, z_i$; i = 1… N$\}$ such that:

(4.13) $$g_A = \frac{g_L}{1-\gamma}$$

(4.14) $$\frac{p_i K_i}{Y_i} = \alpha \frac{1-\tau_{Ki}}{r+\delta}$$

(4.15) $$k_i = h_i (K_i / Y_i)^{\alpha/(1-\alpha)}$$

(4.16) $$\Omega_i (1-\alpha)\lambda k_i (1-a_i) - g_A a_i /(1-a_i) + \varepsilon(1-a_i) = r$$

(4.17) $$a_i = 1 - \frac{g_A}{\lambda s_{Ri} k_i + \varepsilon}$$

(4.18) $$A_i = a_i A_i^*$$

(4.19) $$A_i^* = z_i \tilde{A}$$

(4.20) $$\tilde{A} = (vL)^{1/(1-\gamma)}$$

(4.21) $$v = (1/g_A)\sum_l \lambda k_l s_{Rl} a_l \omega_l$$

(4.22) $$z_i^{(1-\gamma)+1} = (\sigma / v g_A)\sum_l \lambda k_l s_{Rl} a_l \omega_l z_l e^{-\theta d(i,l)}$$

where the last equation comes from (JEK) together with (4.8).

If we knew the relevant parameters and tax rates and wanted to solve for an equilibrium, we would first start by solving for g_A from equation (4.13). Given data for exogenous variables h_i, p_i and τ_{Ki} we could then calculate equilibrium k_i using (4.14) and (4.15). Together with g_A and parameter ε, equation (4.16) would yield a_i. From (4.17) we would then obtain s_{Ri}. Up to this point, there is no interaction across countries, so these results do not depend on geography or θ; this dimension becomes relevant in obtaining actual productivity levels, because they depend on the variables z_i, which capture spillovers from the rest of the world to country *i*. To see how this operates, note that given

the value of θ, equation (4.22) configures a system of N equations (where N is the number of countries) in N unknowns ($z_1, z_2, \ldots z_N$). The solution to this system determines z_i. Given parameter σ, equation (4.20) determines $\tilde{A}$, which together with z_i determines each country's technology frontier A_i^* (equation (4.19)). Finally, from equation (4.18), a country's technology frontier together with its relative *A* level a_i determines A_i.

For the calibration exercise, the first step is to specify the variables we observe and how they relate to the model. We take human capital to be $h_i = e^{\varphi * MYS_i}$, where MYS_i is mean years of schooling of the adult population in country *i*, obtained from Barro and Lee (2000). We use R&D data from Lederman and Sáenz (2003). The 48 countries in our sample are the ones for which there is R&D data for 1995, as well as the necessary TFP and capital intensity variables described in section 3. The first two columns of Table A1 reproduce the values of the R&D investment rate and the value of *A* for the 48 countries in our sample.

For the basic parameters we use the following values: $\varphi = 0.085$, $\alpha = 1/3$, $\delta = 0.08$, $g_L = 0.011$ and $g_A = 0.015$. For the first three, see our discussion in section 3. The last two (the growth rates) were obtained from OECD average growth rates of *L* and *A* for the period 1960-2000.[18] Using (4.13), the values for the two growth rates imply $\gamma = 0.31$. To calculate the net private rate of return, *r*, which we assume to be common across countries, we take the capital income tax in the U.S. to be 25% ($\tau_{K,US} = 0.25$).[19] Given the 1995 U.S. nominal capital-output ratio of 1.5 (see section 3 for how we constructed capital-output ratios), this implies from (4.14) that $r = 8.6\%$. Given this level for *r*, we then use equation (4.14) together with country nominal capital-output ratios to obtain each country's implicit income tax τ_{Ki}.

Remaining parameters we must calibrate are ε, λ, μ and θ. Unfortunately, there is no empirical work that we can rely on to pin down ε. Thus, we choose a value for

[18] Specifically, the growth rate of *A* is the annual growth rate of the weighted average of *A* in the OECD with weights given by employment levels in 1960. OECD membership is defined by 1975 status.
[19] Auerbach (1996) estimates an effective tax rate in the U.S. of about 16%, but King and Fullerton (1984) estimate a much higher level of around 35%. We use 25% as an intermediate value.

ε based on the following reasoning. First, ε cannot be much higher than g_A. This is because for $ks_R \geq 0$ equation (4.17) implies that $a \geq 1 - g_A/\varepsilon$. Thus, a high value of ε would imply that some countries' relative empirical *A* becomes lower than the theoretical minimum $1 - g_A/\varepsilon$. In other words, if free technology diffusion is too important, then it would be hard to account for countries with very low *A* levels. Second, if $\varepsilon < g_A$, then countries with a low value of ks_R ($\lambda s_R k < g_A - \varepsilon$) would not be able to keep up with the world's rate of growth in technology, so they would not have a steady state relative *A* level. (Consistent with stable long run relative income levels, Figure 3 showed roughly parallel slopes for average income across deciles over 1960-2000, with each decile based on 1960 income.) Thus, it seems reasonable to impose the intermediate condition that $\varepsilon = g_A$. We believe, however, that future empirical work should attempt to understand the importance of free technology diffusion captured by parameter ε.

Given this choice for ε, we use two empirical findings to pin down parameters λ and μ, namely that the social rate of return to R&D in the U.S. is three times the net private rate of return (Griliches, 1992) and that the U.S. imposes a subsidy of 20% on R&D (Hall and Van Reenen, 2000), implying that $\tau_{R,US} = -0.2$. Given data for s_R and *k* for the U.S. in 1995 ($s_{RUS} = 2.5\%$ and $k_{US} = 3.6$), then this restriction together with equation (4.17) implies $a_{US} = 0.7$ and $\lambda = 0.38$.[20] From (4.16) we then obtain $\mu = 0.55$.

A parameter remaining to calibrate is θ.[21] Before discussing possible values for this parameter, it is useful to consider the case where $\theta = 0$ – so that there is no effect of distance on international spillovers – and to compare the implications of the model to the data. Using the R&D investment rate data of Lederman and Saenz (2003) and our

[20] Due to the non-linearity of the expression for the social rate of return to R&D, there are actually two values of λ which are compatible with a social rate of return equal to 26% (three times the private rate of return). The higher value of λ, however, would imply a high relative *A* level for the U.S. and consequently – given measured *A* for the U.S. – a value for *A** that would be lower than the measured *A* levels of the high *A* countries, such as Hong Kong and Italy. To avoid this, we choose the lower value of λ.

[21] We must also set a value for σ, which is crucial for determining the level of $\tilde{A}$. We use the value of A_{US} obtained from the data, (4.18)-(4.20), $a_{US} = 0.7$, and a value for z_{US} (equal to one when $\theta = 0$ and a known value from the solution to the above system of equations for the case $\theta > 0$) to arrive at a value for σ.

estimated *k* levels, equation (4.17) yields the model's implied relative *A* level for each country (a_i). We want to compare this against the data. To do so, we use the value of *A* we calculated for the U.S. in the previous section and $a_{US} = 0.7$ to obtain an implied value for the world technology frontier, *A** (recall that with $\theta = 0$ there is a well defined technology frontier that is common to all countries). We can then obtain the model's implied *A* values for all countries using $A_i = a_i A^*$. The result of this exercise is shown in Table 7, where we divide countries into four groups according to their levels of *A* and show the median of the different variables for each group. It is clear that the model does badly for the poorest countries, predicting much lower *A* levels for them than occur in the data. This discrepancy does not occur for the richest countries, so the model is predicting significantly larger *A* differences than in the data. For example, whereas (according to the data) the top group's median *A* is 3.4 times the median *A* of the bottom group, the model implies a ratio of 5.6.

The model implies large differences in productivity in response to small differences in R&D investment rates. As is well known, the neoclassical model – with only around 1/3 share for physical capital – cannot generate large differences in steady state labor productivity in response to modest differences in investment rates (see the discussion in Lucas, 1990). It is worth pausing here to explore some of the reasons behind these divergent properties. Manipulating the neoclassical model, one can show that the semi-elasticity of steady state labor productivity with respect to the investment rate is given by:

$$\frac{\partial \ln y}{\partial s} = \left(\frac{r}{1-\alpha}\right)\left(\frac{1}{\delta + g_A + g_L}\right) \tag{4.23}$$

With the values we used above ($\alpha = 1/3$, $\delta = 0.08$, $g_L = 0.011$ and $g_A = 0.015$), (4.23) yields a semi-elasticity of only 1.22% when evaluated at $r = 8.6\%$. Thus large differences in investment rates would be required to generate sizable differences in labor productivity across countries. Two differences between the way the R&D investment rate operates in our model and the way the physical capital investment rate operates in the neoclassical model stand out: first, the depreciation rate of ideas in our model is zero versus $\delta = 0.08$ for capital in the neoclassical model; second, the elasticity of output with respect to the stock of ideas can exceed 1/3 (we have it at 2/3). To see the importance of these values,

note that with $\alpha = 2/3$ the semi-elasticity doubles to 2.46% (still with $r = 8.6\%$). If we use $\delta = 0$ as well, then the semi-elasticity increases to 9.6%. In our model, the combined share of physical capital and ideas is actually 1. Without the constraint of the world technology frontier, therefore, the long run response of output would be infinite.

It is important to recall that the results shown in Table 7 and discussed above were derived for the case of $\theta = 0$. Is it possible that a positive value of θ could improve the model's fit with the data? As will become clear below, countries with high levels of *k* and high R&D investment rates tend to cluster together. Thus, assuming a positive value for θ would actually make the model *less* consistent with the data, since it would imply an even larger difference between *A* levels across rich and poor countries.

One possible reason why the model is not doing well in matching the data is that measured R&D is not the appropriate empirical counterpart of "research" in the type of models we have been examining. In particular, measured R&D only includes formal research; this is research performed in an R&D department of a corporation or other institution. This fails to capture informal research, which may be particularly important in non-OECD countries. To explore this idea, in the rest of this section we assume that both R&D intensity and the productivity index *A* are measured with error. We estimate "true" R&D intensities by minimizing a loss function equal to the sum of two terms that capture, respectively, the deviation of the "true" R&D intensities from the data and the deviation of the model's implied (log of) *A* values from the data, with weights given by the standard deviation of the corresponding differences.[22] In principle, we could follow this procedure for each value of θ. However, at $\theta = 0$, the partial derivative of our loss function with respect to θ is positive and large, implying that – just as argued above – the model's fit with the data worsens as θ increases from zero. Thus, we restrict ourselves to estimating R&D intensities for $\theta = 0$ and later show what happens if, keeping the same R&D intensities estimated for $\theta = 0$, we have positive values of θ.

[22] We do this in two stages. In the first stage we minimize a loss function without weights. We use the results to calculate the standard deviation of the error terms, or differences between data and "true" values for both R&D intensity and productivity. In the second stage we minimize the loss function with weights given by these calculated standard deviations.

It should be acknowledged that this procedure obviously implies that we can no longer evaluate the model's consistency with the data; our interest is now to explore the implications of the model for the differences in R&D investment rates that would be necessary to explain cross-country differences in *A*, as well as the implied differences in R&D tax rates that would be necessary to bring about those R&D investment rates.

The results of the exercise described above are shown in Figure 4 and Table A1 (columns 3 and 4). There are three points to note from these results. First, it is clear that the procedure leads to only small deviations of *A* from the data, whereas the deviations are more significant for R&D intensities. It would appear that R&D intensities have more significant measurement problems (or are conceptually more different than research intensity in our model) than productivity levels. Indeed, the standard deviation of residuals of s_R with respect to the data is 0.12, whereas the corresponding value for the (log of) *A* is 0.01.[23] Second, there are some countries for which the estimated R&D intensity is much higher than the data. Italy, for example, has a measured R&D intensity of 1.1%, whereas its "true" value is 8.3%. This arises because of Italy's high measured productivity (Italy's *A* is 24% higher than the U.S. level) and low value of *k* (2.6 versus 3.6 in the U.S.). Something similar happens for other high-*A* countries, such as Hong Kong and Ireland. Finally, just as one would expect given the results above, estimated R&D intensities vary much less than the corresponding values in the data. This is the main mechanism by which the procedure allows the model to fit perfectly. It also suggests that measurement error may be behind the low R&D intensities of several poor countries and of some high *A* countries such as Italy, Ireland and Hong Kong.

We can now explore what happens when θ is positive, so that spillovers decline with distance. Given the estimated R&D intensities, productivity levels change with θ only because of the associated changes in the variables z, which capture the effect of distance on spillovers for each country. In principle, we can obtain the values of $(z_1, z_2, \ldots z_{48})$ for any $\theta \geq 0$ from the solution of a system of 48 non-linear equations represented by (4.22). Equation *i* of this system can be expressed as:

[23] These standard deviations are the ones that arise after the two stage procedure described in the previous footnote. After the first stage, the standard deviations for the R&D rate and the (log of) *A* are 0.11 and 0.03, respectively.

$$z_i^{(1-\gamma)+1} = \frac{\sum_{j=1}^{48} \lambda k_l s_{Rl} a_l z_l \omega_l e^{-\theta d(i,l)}}{\sum_{l=1}^{48} \lambda k_l s_{Rl} a_l \omega_l} \tag{4.24}$$

Solving this system numerically for the parameter values we have discussed and the R&D intensities derived before, we arrive at a value of z_i for each country, from which we can then obtain the country's level of *A* by using $A_i = a_i z_i \tilde{A}$ from (4.18) and (4.19).

What are reasonable values to use for the parameter θ? Using industry level data on productivity and research spending across the G-5 countries, Keller (2002) estimated a reduced form model where cumulative industry research affects own productivity and also affects productivity in the same industry in other countries through international spillovers that decline with distance.[24] Given the similarity between Keller's system and a reduced form of our model, it seems reasonable to use Keller's estimate of θ, namely $\theta_K \equiv 0.0009$ in the calibration of our model. It turns out, however, that with $\theta = \theta_K$ our model cannot match the data – in particular, there is no solution to the system of equations (4.24), at least for the parameters used for the exercises above. This is because θ_K is unreasonably high. One way to see this is by noting that it implies a half distance of 746 miles: this implies that spillovers from the U.S. to Japan would be only one tenth of those to Mexico, and spillovers from the U.S. to New Zealand would be only one fifth of those to Japan.

We were able to find solutions for the system with $\theta = \theta_K / 5$. For comparison, we also obtained solutions for two other values of θ, namely $\theta = \theta_K / 10$ and $\theta = \theta_K / 100$. A group of European countries (Belgium, France, United Kingdom, Germany, Ireland, Italy, and Netherlands) always come out with the highest values of z, whereas New Zealand always comes out with the lowest value. For $\theta = \theta_K / 100$, $\theta = \theta_K / 10$ and $\theta = \theta_K / 5$, the minimum and maximum values of z are (93%, 96%), (48%, 68%) and (24%, 50%), respectively. Clearly, for high values of θ, geography by itself can lead to large differences in productivity across countries.

[24] For other estimates of international spillovers from R&D, see Coe and Helpman (1995) and Coe, Helpman and Hoffmaister (1997). For a study of agricultural R&D spillovers, see Evenson and Gollin (2003). Becker, Philipson and Soares (2003) present evidence consistent with international spillovers of health technology.

In the rest of this section, we focus on the case $\theta = 0$, since – as explained above – the model's fit with the data is best at this point. (Recall that the model fits perfectly because we are using the estimated research intensities and the implied *A* values). Table 8 presents summary statistics for the solution for the case of $\theta = 0$. Our discussion of these results will focus on the comparison of the poorest and richest quartiles (ordered, as above, in terms of *A* levels) in this table.

There are several points that we want to highlight in relation to these results. First, the median income tax is 13% and 6% for the poorest and richest countries, respectively. Everything else equal, this would lead to a lower R&D investment rate in the poorest countries. Second, as expected, rich countries have a higher *k* than poor countries: the level of *k* in these two groups is 2 and 2.9, respectively. As commented in Section 4B, higher *k* has a direct effect on relative *A* (see equation (4.17)) and an indirect effect (it could be positive or negative) through its impact on R&D investment rates (see equation (4.16)). A natural question arises: is it the case that once we take into account the effect of *k* on TFP we can resuscitate the "neoclassical revolution" mantra that differences in physical and human capital accumulation rates account for most of cross-country income differences? More concretely, how much of the variation in *A* levels across countries is due to the variation in levels of *k*? A simple way to answer this question is to note from equation (4.17) that differences in relative *A* levels are driven by differences in the product $s_R k$ across countries. Running a regression of s_R on the log of this product yields a coefficient of 0.8, which implies that when $s_R k$ increases by one percent, we should expect s_R to increase by 0.8%. Clearly, most of the variance of the product $s_R k$ is accounted for by the variance of s_R.

Third, the social return to R&D is higher for poor countries. This is consistent with the findings in Lederman and Maloney (2003) and also with the idea that poor countries have policies and institutions that negatively affect the quantity of research.

Fourth, the column with heading τ_R indicates the R&D tax rate required to produce the "true" R&D investment rates given each country's levels of τ_K. The main question we address here is whether differences in income tax rates, which affect both the rate of

investment in physical capital and R&D, are sufficient to explain differences in estimated research intensities. The answer is clearly negative: the required R&D tax rate in the poorest countries is 102% compared to -16% in the richest countries. To address the same question from a different angle, the last column calculates each country's implied relative *A* level if all countries had the same R&D tax as the U.S. but kept their own levels of τ_K. It is clear that differences in τ_K alone are too small to account for the wide dispersion in productivity levels across countries.

Finally, as emphasized above, the results in Table 8 suggest that small differences in steady-state R&D investment rates have large effects on steady state relative *A* levels. For example, in the calibrated model, by increasing its R&D investment rate by 1% from 0.6% India could double its steady state relative *A* level from 17% to 34%, clearly a very large effect. India's social rate of return to research, however, is a moderate 30%. The apparent contradiction arises because the large effect of the increase in the R&D intensity on the relative *A* level is a steady-state comparative-statics result, and hence does not take into account the transition, which is a crucial component in the calculation of the social rate of return to R&D. As a result, in spite of the large effect of differences in R&D investment rates on relative *A* levels in steady state, the required implicit taxes on R&D are not huge.

4E. The benefits of engagement

One of the benefits of the model we have constructed is that it allows us to perform an interesting exercise. We can ask: how much do countries benefit from spillovers from the rest of the world?

First, note that a country's equilibrium a_i is not affected by being isolated or engaged. Thus, the whole benefit of engagement is going to captured by the way engagement affects the term z_i. Now, if a country is isolated, or disengaged, its equilibrium z would be characterized by the solution to the system (4.24) when $\theta \to \infty$. It is easy to check that this yields

(4.25) $$\breve{z}_i = \left(\frac{\lambda k_i s_{Ri} a_i \omega_i}{\sum_l \lambda k_l s_{Rl} a_l \omega_l} \right)^{1/(1-\gamma)}$$

Thus, the benefits of engagement are captured by $z_i / \breve{z}_i$. From (4.17) we get

$$z_i / \breve{z}_i = z_i \left(\frac{\sum_l \omega_l \upsilon_l}{\omega_i \upsilon_i} \right)^{1/(1-\gamma)}$$

where $\upsilon_i \equiv a_i \lambda k_i s_{Ri} = \lambda R_i / A_i^* L_i$ is a measure of research intensity. Letting $\bar{\upsilon} \equiv \sum_j \omega_j \upsilon_j$ be the world's weighted average of υ_i, we obtain

(4.26) $$\frac{z_i}{\breve{z}_i} = z_i \left(\frac{1}{\omega_i} \right)^{1/(1-\gamma)} \left(\frac{\bar{\upsilon}}{\upsilon_i} \right)^{1/(1-\gamma)}$$

The first term on the RHS of this equation, z_i, captures the fact that even when fully engaged, a country's technology frontier is inferior to the world's frictionless frontier if $\theta > 0$, in which case $z_i < 1$ for all i. The second term is the pure scale effect that arises in this model. The third term, which we call the "Silicon Valley" effect, captures the fact that richer countries benefit less from being part of the world than poor countries do because of their higher effective research intensity.

Table 9 presents results based on these values and assuming $\theta = 0$, which implies $z_i = 1$ for all i. The results suggest huge benefits of engagement. At the extreme, Senegal's productivity is 187 thousand times higher than it would be if it was isolated! Of course, if $\theta > 0$ then $z_i < 1$ and the overall effect would be small. Still, it is our conjecture that any reasonable value of θ would still imply enormous benefits of engagement. Of course, in a more general model, it is reasonable to think that productivity could not fall below a certain level because of Malthusian forces. Specifically, suppose there is a fixed factor such as land. Then, for sufficiently low *A*, population would decline until income per capita was equal to the subsistence level. Instead of very low levels of *A*, disengagement would mean

very low population sizes. Put differently, an important part of the benefits of engagement may be realized through larger population rather than higher productivity. The implications are clear: if it were not for the benefits of sharing knowledge internationally, countries would have much lower productivity levels and populations than they now do.

4F. Discussion of main results

We finish this section with a discussion of the main results we want to emphasize.

First, the usual separation between capital and productivity – or between investment and technological change – is not always valid. For a given R&D investment rate, higher investment rates in physical and human capital lead naturally to higher TFP productivity levels. Thus, one should not jump from cross-country dispersion in TFP to the conclusion that differences in physical and human capital play a minority role in accounting for international income differences. When we calibrate our model, however, we find that differences in R&D investment rates account for most of the cross country variation in productivity.

Second, international variation in R&D investment appears *more* than large enough to generate the international variation in productivity. But it seems likely that measured R&D does not capture all of the investment associated with adoption of foreign technology. Indeed, we find that countries such as Indonesia, Peru and Senegal have R&D investment rates that are much too low to be consistent with their productivity levels. It is likely that their true research intensities are much higher than the measured ones. We hope to see more research in understanding how to capture and measure "research".

Third, differences in (implicit) capital income tax rates are not large enough to account for the observed differences in R&D investment rates and productivity levels. The calibrated model suggests that sizable differences in R&D taxes are needed. These R&D taxes are clearly not formal or explicit taxes, but the result of policies and institutions that make research more costly or reduce its associated returns. Exploring the nature and source of these differences in implicit R&D taxes across countries is an important topic for future research.

Finally and most importantly, the calibrated model indicates that countries benefit enormously from international knowledge spillovers. We think any reasonable value of θ (which governs the rate at which spillovers decline with distance) would yield results similar to those we presented above.

5. Conclusion

Externalities are not theoretically necessary to sustain growth. But they appear essential for understanding why many countries grow at similar rates despite differing investment rates. A dramatic way to summarize the importance of international knowledge externalities is to calculate world GDP in the absence of such externalities. According to our calibrated model, world GDP would be only 6% of its current level, or on the order of $3 trillion rather than $50 trillion, if countries did not share ideas. Such scale effects from the nonrivalry of knowledge are a central theme in the works of Romer (1990), Kremer (1993), Diamond (1997), Jones (2001, 2004) and many others.

Because diffusion is not costless, however, differences in knowledge investments may explain a significant portion of income differences across countries. We show that modest barriers to technology adoption could account for differences in TFP of a factor of four or more, as observed in the data. But we have not documented such barriers to knowledge diffusion in practice. We consider this a priority for future research.

We have also left for future research the identification of the primary channels of international knowledge spillovers. Trade, joint ventures, FDI, migration of key personnel, and imitation may all play important roles. See Keller (2004) for a survey of recent empirical work on this topic. A model with trade would lead naturally to some countries having a comparative advantage in doing innovative R&D and other countries focusing on adoption and imitation R&D. The evidence on international patenting supports the notion that innovative R&D is concentrated in rich countries. Of course, countries can imitate other imitators as well as the original innovators. We hope to see future research documenting not only the vehicles for knowledge diffusion, but their specific routes.

Appendix

The firm's maximization problem can be restated as choosing $\dot{A}_{jis}$ and $\dot{K}_{jis}$ to maximize:

$$\int_t^{\infty}\left((1-\tau_{Ki})K_{jis}^{\alpha}(A_{jis}h_i)^{1-\alpha}-p_i\dot{K}_{jis}-p_i\delta K_{jis}-\frac{(1+\tau_{Ri})}{(1-\mu)\lambda}\left(\frac{\dot{A}_{jis}}{1-A_{jis}/A_{is}^{*}}-\varepsilon A_{jis}-\mu\lambda\bar{R}_{is}\right)\right)e^{-r(s-t)}ds$$

Letting Q represent the expression in the integral, then we know that a solution to this problem must satisfy the following Euler Equations: $\partial Q/\partial K_{jis}=\frac{d}{ds}(\partial Q/\partial\dot{K}_{jis})$ and $\partial Q/\partial A_{jis}=\frac{d}{ds}(\partial Q/\partial\dot{A}_{jis})$. The first Euler equation is:

$$\frac{\alpha Y_{jis}}{p_i K_{jis}}=\frac{r+\delta}{1-\tau_{Ki}}$$

Since in a symmetric equilibrium the capital-output ratio of firm j is the same as the aggregate capital output ratio, then this implies that:

$$\frac{p_i K_{it}}{Y_{it}}=\alpha\frac{1-\tau_{Ki}}{r+\delta}$$

As to the second Euler equation, differentiation yields (we are using the symmetry condition for the equilibrium):

$$\frac{\partial Q}{\partial A_{jis}}=\left((1-\tau_{Ki})(1-\alpha)Y_{jis}/A_{jis}-\frac{(1+\tau_{Ki})}{(1-\mu)\lambda}\left(\frac{g_A a_i}{(1-a_i)^2}-\varepsilon\right)\right)e^{-r(s-t)}$$

and

$$\partial Q/\partial\dot{A}_{jis}=-\frac{(1+\tau_{Ri})}{(1-\mu)\lambda(1-a_i)}e^{-r(s-t)}$$

Hence,

$$\frac{d}{ds}(\partial Q/\partial\dot{A}_{jis})=\frac{r(1+\tau_{Ri})}{(1-\mu)\lambda(1-a_i)}e^{-r(s-t)}$$

Thus, the Euler equation is:

$$(1-\tau_{Ki})(1-\alpha)Y_{jis}/A_{jis}-\frac{(1+\tau_{Ri})}{(1-\mu)\lambda}\left(\frac{g_A a_i}{(1-a_i)^2}-\varepsilon\right)=\frac{r(1+\tau_{Ri})}{(1-\mu)\lambda(1-a_i)}$$

Noting that in a symmetric equilibrium we must have $Y_{jis}/A_{jis}=Y_{is}/A_{is}L_{is}=k_i$, and manipulating, we get:

$$\Omega_i(1-\alpha)\lambda k_i(1-a_i)-g_A a_i/(1-a_i)+\varepsilon(1-a_i)=r$$

where $\Omega_i \equiv \frac{(1-\tau_{Ki})(1-\mu)}{(1+\tau_{Ri})}$.

Comparative statics

(From here onwards we drop the subscripts). It is easy to show that a is increasing in both Ω and k. In particular:

$$\frac{\partial a}{\partial k} = \frac{\Omega(1-\alpha)\lambda(1-a)}{\Omega(1-\alpha)\lambda k + \left(g_A/(1-a)\right)(1+1/(1-a)) + \varepsilon} > 0$$

Differentiating $g_A = (\lambda k s + \varepsilon)(1-a)$ (using s for s_R) we get

$$(\lambda s dk + \lambda k ds)(1-a) - da(\lambda k s + \varepsilon) = 0$$

This implies that:

$$\lambda k(\partial s/\partial k) = \frac{(\partial a/\partial k)(\lambda k s + \varepsilon)}{(1-a)} - \lambda s$$

Plugging in from the result above we finally get:

$$k(\partial s/\partial k) = \frac{\Omega(1-\alpha)(\lambda k s + \varepsilon)}{\Omega(1-\alpha)\lambda k + \left(g_A/(1-a)\right)(1+1/(1-a)) + \varepsilon} - s$$

Summing on the RHS and noting that the denominator is clearly positive we get that $\partial s/\partial k > 0$ if and only if:

$$\Omega(1-\alpha)\varepsilon - s(g_A/(1-a))(1+1/(1-a)) - \varepsilon s$$

This could well be negative!

Table 1

Some Growth Models by Type of Externality

	New Good Externalities	**No New Good Externalities**
Knowledge Externalities	Stokey 1988 & 1991 Romer 1990 Aghion and Howitt 1992 Eaton and Kortum 1996 Howitt 1999 & 2000	Romer 1986 Lucas 1988 & 2004 Tamura 1991 Parente and Prescott 1994
No Knowledge Externalities	Rivera-Batiz and Romer 1991 Romer 1994 Kortum 1997	Jones and Manuelli 1990 Rebelo 1991 Acemoglu and Ventura 2002

Table 2

Output Growth Declined Sharply Worldwide

	Average *Y/L* Growth			Average s_I			Average s_H		
	1960-75	1975-00	# of countries	1960-75	1975-00	# of countries	1960-75	1975-00	# of countries
World	2.7%	1.1%	96	15.8%	15.5%	96	7.1%	9.7%	74
OECD	3.4	1.8	23	23.2	22.9	23	11.4	14.3	21
Non-OECD	2.5	0.9	73	13.5	13.2	73	5.4	8.0	53
Africa	2.0	0.5	38	12.3	10.5	38	3.9	6.0	19
Asia	3.2	2.8	17	14.5	19.9	17	6.9	9.9	16
Europe	3.8	1.9	18	24.9	23.1	18	10.7	13.7	16
North America	2.8	0.4	13	14.3	14.5	13	7.5	10.2	13
South America	2.3	-0.1	10	17.3	15.0	10	7.1	9.8	10
1st quartile (poorest)	1.6	0.5	24	9.6	9.9	24	3.1	5.0	19
2nd quartile	2.6	1.4	24	14.8	14.2	24	5.7	8.9	19
3rd quartile	3.5	1.1	24	15.4	16.3	24	7.5	10.3	18
4th quartile (richest)	3.0	1.5	24	23.6	21.9	24	12.3	15.1	18

Notes: *Y/L* is GDP per worker. s_I is the physical capital investment rate, and s_H years of schooling attainment (for the 25+ population) divided by 60 years (working life). Data Sources: Barro and Lee (2000) and Heston, Summers, and Aten (2002).

Table 3

Investment Rates Are More Persistent than Growth Rates

	1980-2000 vs. 1960-1980			Decade to Decade		
	***Y/L* Growth**	S_I	S_H	***Y/L* Growth**	S_I	S_H
World	**.34** (.13)	**.56** (.07)	**1.02** (.04)	**.20** (.07)	**.77** (.04)	**1.00** (.02)
OECD	.12 (.13)	**.44** (.09)	**.86** (.08)	**.27** (.09)	**.70** (.06)	**.92** (.03)
Non-OECD	**.36** (.17)	**.44** (.09)	**1.10** (.07)	**.17** (.08)	**.71** (.05)	**1.04** (.03)

Notes: World = 74 countries with available data; OECD = 22 countries; and non-OECD = 52 countries. Decades consisted of the 1960s, 1970s, 1980s, and 1990s. All variables are averages over the indicated periods. Each entry is from a single regression. Bold entries indicate p-values of 1% or less. Data Sources: Barro and Lee (2000) and Penn World Table 6.1 (Heston, Summers and Aten, 2002).

Table 4

Investment Rates Correlate More with Levels than with Growth Rates

	Independent Variable = S_I			**Independent Variable = S_H**		
	Dependent Variable			Dependent Variable		
	Y/L Growth Rates	*Y/L* Log Levels	# of countries	*Y/L* Growth Rates	*Y/L* Log Levels	# of countries
All countries	**.111** (.017) $R^2 = .32$	**1.25** (0.13) $R^2 = .48$	96	**.210** (.060) $R^2 = .15$	**.313** (.026) $R^2 = .67$	74
OECD	.020 (.047) $R^2 = .01$	.760 (.358) $R^2 = .18$	23	**-.259** (.078) $R^2 = .37$	**.119** (.024) $R^2 = .56$	21
Non-OECD	**.124** (.023) $R^2 = .29$	**.842** (.162) $R^2 = .28$	73	**.367** (.095) $R^2 = .22$	**.314** (.043) $R^2 = .51$	53

Notes: Variables are averages over 1960-2000. Each entry is from a single regression. Bold entries indicate p-values of 1% or less. Data Sources: Barro and Lee (2000) and Penn World Table 6.1 (Heston, Summers and Aten, 2002).

Table 5

R&D Intensity Also Correlates More with Levels than Growth Rates

Independent Variable = R&D Spending as a Share of GDP

	Dependent Variable			Dependent Variable		
	Y/L Growth Rates	*Y/L* Log Levels	# of countries	*TFP* Growth Rates	*TFP* Log Levels	# of countries
All countries	0.40	**0.69**	82	0.43	**0.37**	67
	(0.59)	(0.23)		(0.52)	(0.08)	
	R^2 = .01	R^2 = .10		R^2 = .01	R^2 = .27	
OECD	-0.15	**0.42**	21	-0.16	**0.17**	21
	(0.46)	(0.11)		(0.32)	(0.06)	
	R^2 = .01	R^2 = .45		R^2 = .01	R^2 = .28	
non-OECD	0.88	0.55	61	0.85	**0.34**	46
	(1.03)	(0.41)		(1.01)	(0.14)	
	R^2 = .01	R^2 = .03		R^2 = .02	R^2 = .12	

Notes: Variables are country averages over years in 1960-2000 with data relative to time effects. *Y/L* is GDP per worker. *TFP* nets out contributions from human and physical capital, as described in the text. Each entry is from a single regression. Bold entries indicate p-values of 2% or less. Data Sources: Barro and Lee (2000), Penn World Table 6.1 (Heston, Summers and Aten, 2002), and Lederman and Saenz (2003).

Table 6

Alternative Ways of Modeling International Spillovers

	Spillovers	**Growth rate**	**Advantages**	**Disadvantages**
H1	$\dot{A}^* = \sigma \sum_i \left(\frac{\lambda R_i}{L_i} \right)$	$g_A = \sigma \sum_i \lambda k_i s_{Ri} a_i$	No scale effects	Larger countries contribute no more to g_A than do small countries
H2	$\dot{A}^* = \sigma \sum_i \left(\frac{\lambda R_i}{L} \right)$	$g_A = \sigma \sum_i \lambda k_i s_{Ri} a_i \omega_i$ where $\omega_i = L_i / L$	Previous ones plus: Size matters for a country's contribution to g_A	Countries with higher than average $k_i s_{Ri} a_i$ would be better off ignoring research from the rest of the world
J	$\dot{A}^* = \left(A^* \right)^{\gamma - 1} \sigma \sum_i \lambda R_i$	$g_A = g_L / (1 - \gamma)$	Previous ones plus: Research-intensive countries do not prefer to disengage from the rest of the world.	g_A does not depend on R&D efforts…but is this a disadvantage? (See Jones, 1995)
JEK	$\dot{A}_i^* = \left(A_i^* \right)^{\gamma - 1} \sigma \sum_l \lambda R_l \eta_{il}$	$g_A = g_L / (1 - \gamma)$	Previous ones plus: The model takes into account effect of distance on spillovers.	We will find it hard to see the cost of geographic isolation in the TFP data.

Table 7

Model *A* versus data *A* ($\theta = 0$ case)

Country	Data k	Data s_R	Data *A*	Model *A*
Quartile 1	2.0	0.4%	4,478	2,184
Quartile 2	2.5	0.5%	9,574	5,358
Quartile 3	3.1	1.7%	11,111	11,763
Quartile 4	2.9	1.7%	15,441	12,286

Table 8

Implied R&D tax rates

	"True"						a for
Country	τ_K	k	s_R	a	**SRR**	τ_R	$\tau_{Ri} = \tau_{R,US}$
Quartile 1	**13%**	**2.0**	**0.60%**	**20%**	**42%**	**102%**	**58%**
Quartile 2	**0%**	**2.5**	**1.13%**	**43%**	**37%**	**93%**	**68%**
Quartile 3	**4%**	**3.1**	**1.97%**	**50%**	**29%**	**31%**	**72%**
Quartile 4	**6%**	**2.9**	**2.98%**	**70%**	**21%**	**-16%**	**70%**

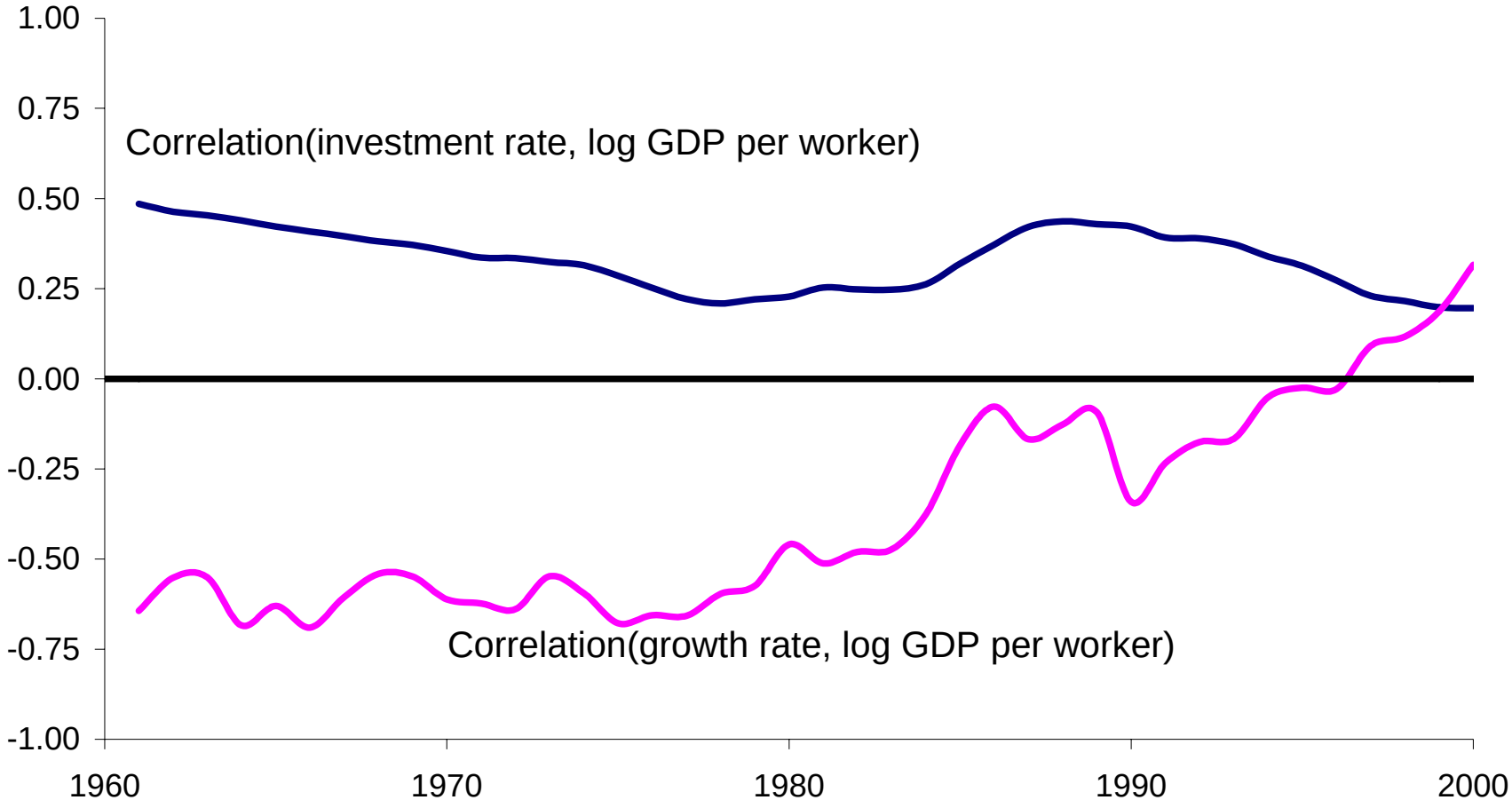
Notes: τ_R is calculated as the level of τ_R needed to generate the "true" research intensity. For each country, we use its own implied income tax level (τ_K) and its own capital intensity level k. The last column presents the equilibrium steady state relative *A* level (a) for the hypothetical case in which all countries have the same R&D tax as the U.S. ($\tau_{Ri} = \tau_{R,US}$) but have different income tax rates and capital intensity levels.

Table 9

Benefits of Engagement for Selected Countries

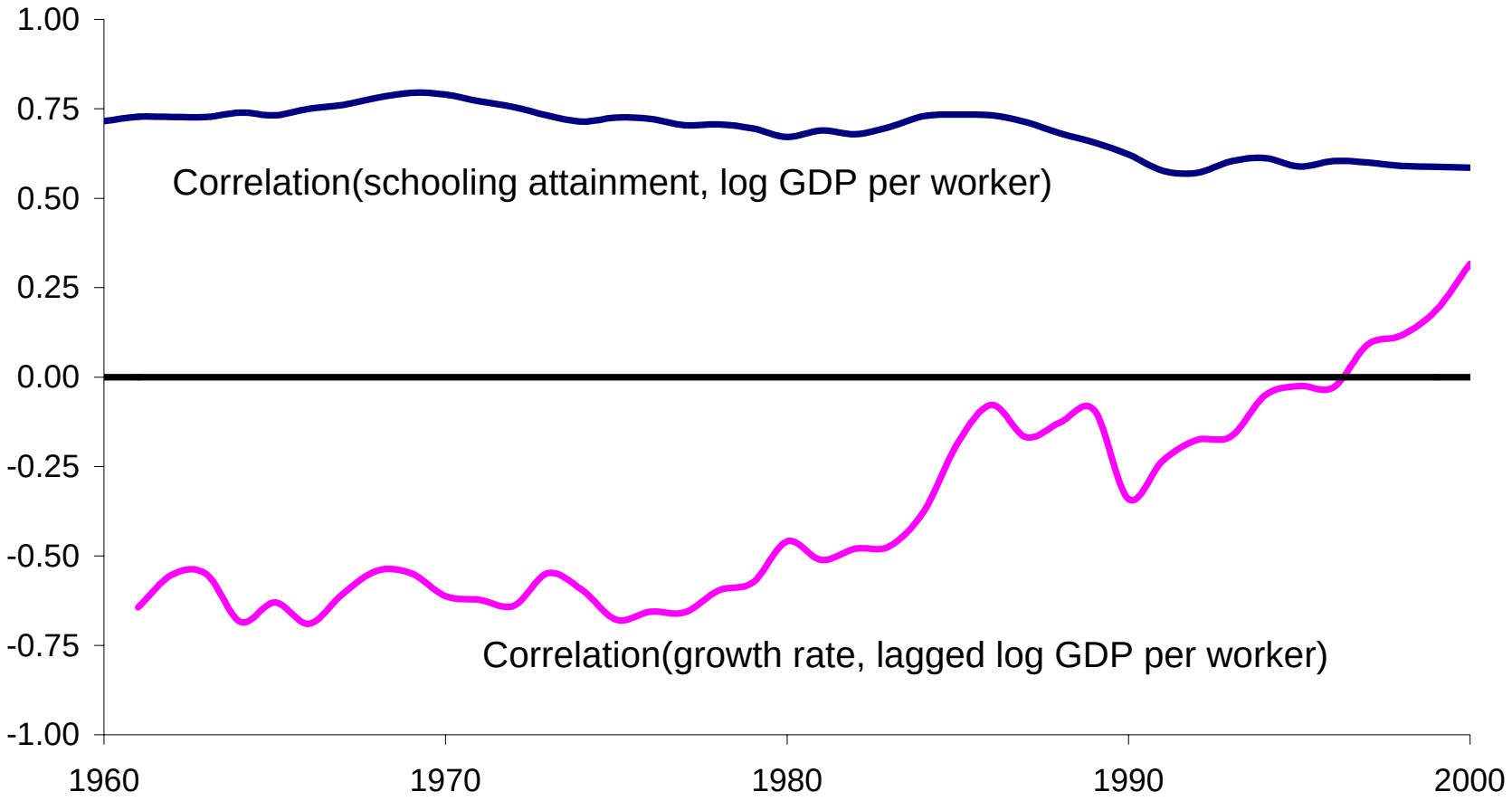
Country	Share of world's *L*	Scale Effect	S.V. Effect	Total Effect
U.S.	**7.1%**	**37**	**0.12**	**5**
U.K.	**1.5%**	**297**	**0.21**	**64**
Belgium	**0.2%**	**4,093**	**0.12**	**480**
Brazil	**3.1%**	**114**	**0.97**	**110**
India	**1.3%**	**9**	**23.0**	**217**
China	**38.7%**	**4**	**70.6**	**258**
Senegal	**0.2%**	**4,451**	**42.0**	**187,035**

Figure 1: OECD Incomes Correlate Negatively with Growth Rates, Positively with Investment Rates



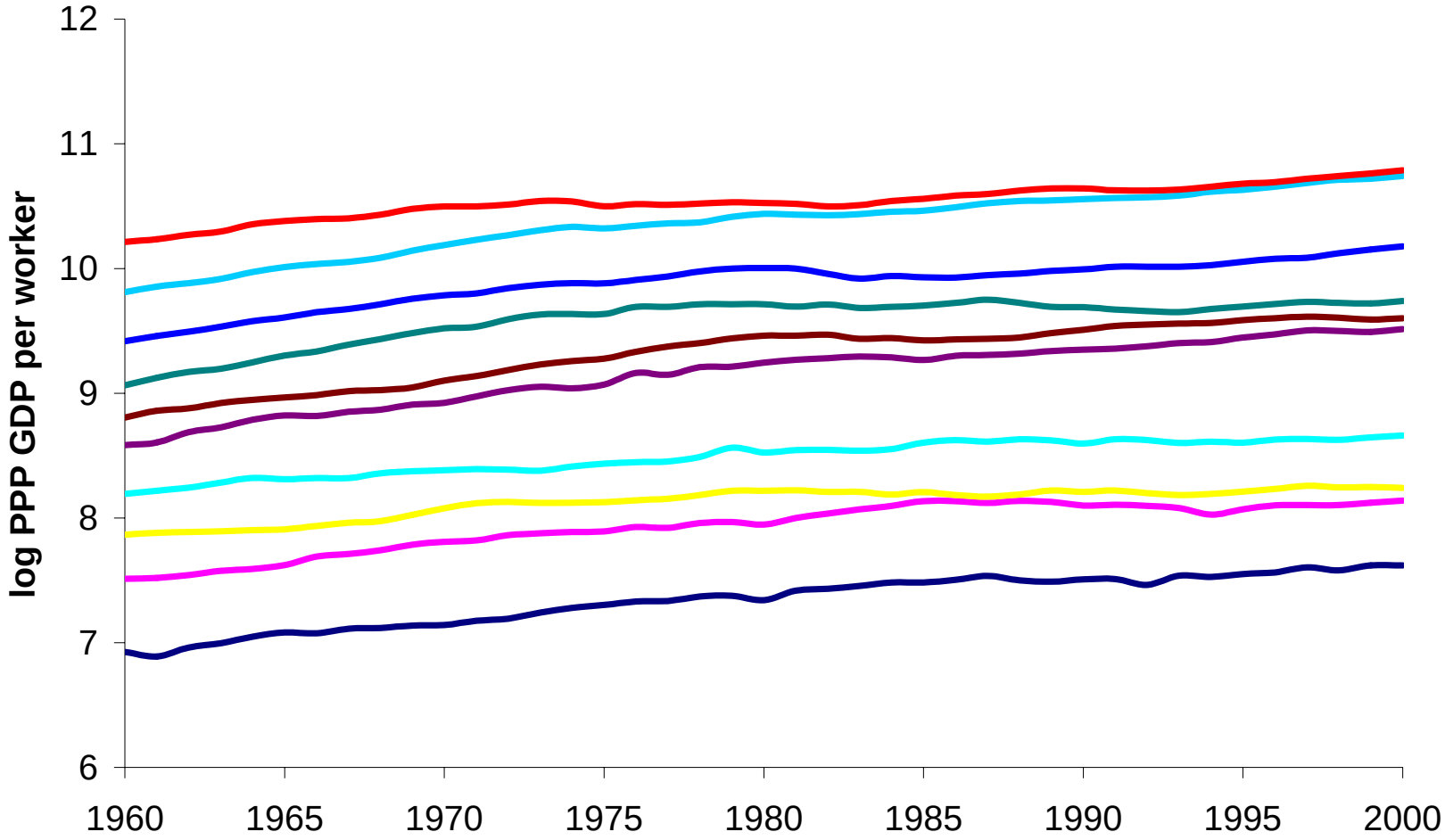
Source: Penn World Table 6.1 data on 23 OECD Countries

Figure 2: OECD Incomes Correlate Negatively with Growth Rates, Positively with Schooling



Sources: Penn World Table 6.1 and Barro and Lee (2000) data for 21 OECD countries.

Figure 3: The Evolution of Income for 1960 Deciles



Source: Penn World Table 6.1.

Figure 4: Deviations of the model from the data for research intensity and productivity

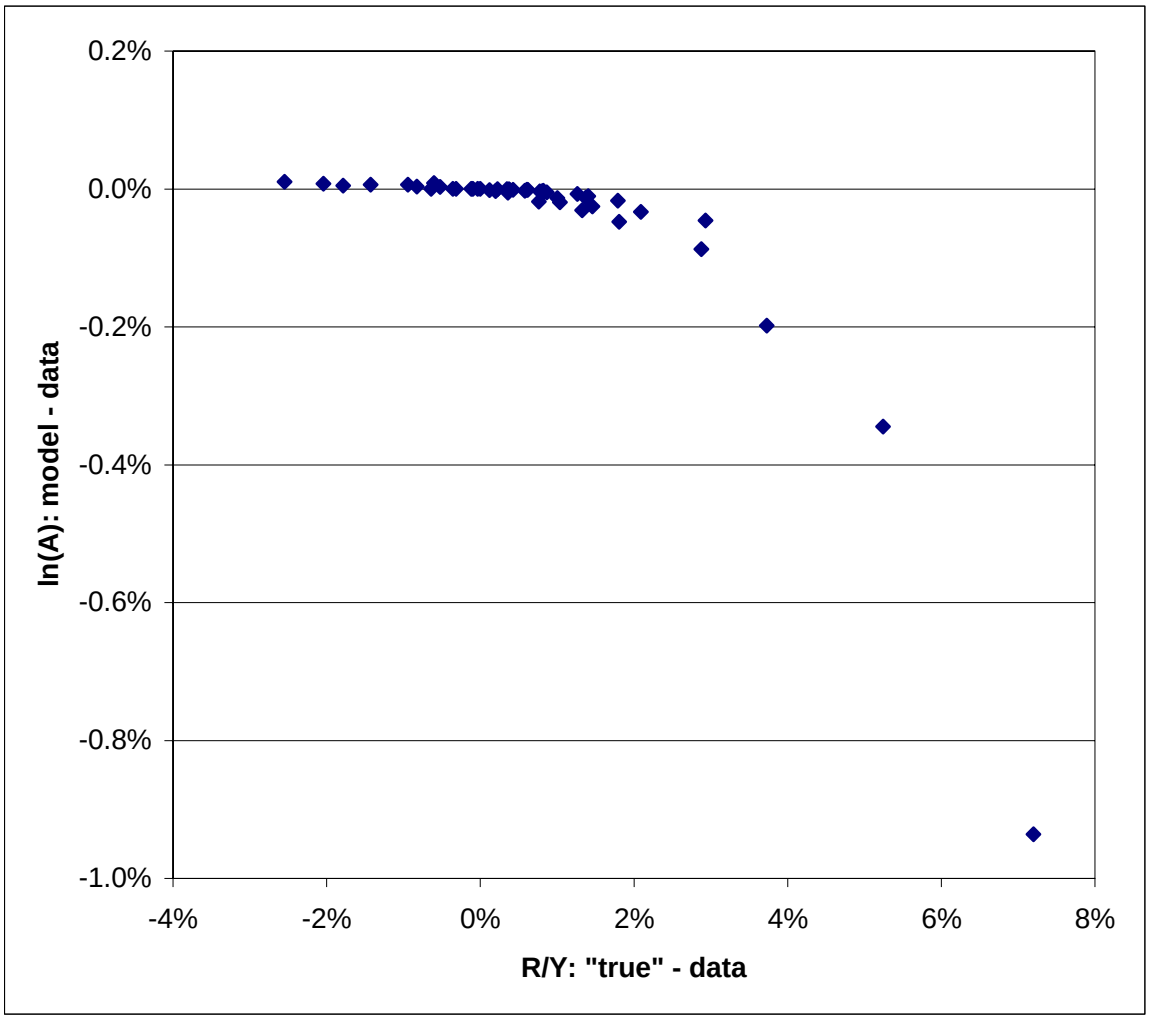


Table A1: Data and "true" values for research intensity and productivity

Country	Data sR	Data *A*	"True" sR	Implied *A*
Argentina	0.41%	9,720	1.21%	9,719
Bolivia	0.37%	4,672	0.74%	4,672
Brazil	0.86%	9,836	1.67%	9,835
Chile	0.61%	11,078	1.98%	11,075
China	0.60%	2,570	0.28%	2,570
Colombia	0.28%	8,143	1.54%	8,141
Ecuador	0.08%	5,990	0.69%	5,990
Egypt	2.11%	11,126	3.57%	11,119
Hong Kong	0.25%	17,874	5.49%	17,732
Hungary	0.73%	7,172	0.63%	7,172
Indonesia	0.09%	5,912	0.91%	5,911
India	0.63%	3,755	0.60%	3,755
Israel	2.75%	13,919	2.15%	13,922
South Korea	2.49%	8,842	0.71%	8,843
Mexico	0.31%	8,781	1.08%	8,780
Panama	0.38%	6,106	0.60%	6,106
Peru	0.05%	4,285	0.40%	4,285
Poland	0.69%	4,893	0.33%	4,893
Romania	0.80%	2,757	0.16%	2,757
Senegal	0.02%	3,069	0.64%	3,068
Singapore	1.16%	13,592	2.16%	13,587
El Salvador	0.33%	11,096	3.26%	11,084
Thailand	0.12%	5,212	0.49%	5,212
Tunisia	0.32%	10,323	2.11%	10,319
Taiwan	1.78%	14,944	3.59%	14,928
Uganda	0.59%	2,878	1.02%	2,878
Uruguay	0.28%	10,088	1.69%	10,085
Venezuela	0.48%	9,427	1.35%	9,426
Austria	1.56%	14,807	2.60%	14,800
Belgium	1.57%	15,597	2.89%	15,586
Canada	1.64%	11,614	1.12%	11,615
Denmark	1.84%	13,678	1.95%	13,677
Spain	0.81%	15,758	3.69%	15,726
Finland	2.37%	10,358	0.94%	10,360
France	2.31%	15,411	3.07%	15,404
United Kingdom	1.99%	13,954	2.35%	13,952
Germany	2.25%	11,993	1.31%	11,994
Greece	0.49%	10,046	1.07%	10,046
Ireland	1.35%	17,177	5.08%	17,098
Italy	1.08%	19,204	8.27%	18,795
Japan	2.89%	9,864	0.85%	9,865
Netherlands	1.99%	14,136	2.19%	14,135
Norway	1.71%	10,990	0.88%	10,991
New Zealand	0.97%	9,911	0.85%	9,911
Portugal	0.57%	13,230	2.65%	13,220
Sweden	3.46%	10,416	0.91%	10,418
Turkey	0.38%	7,800	1.18%	7,800
USA	2.51%	15,472	2.51%	15,472

References

Acemoglu, D. and J. Ventura (2002), "The World Income Distribution," Quarterly Journal of Economics 117: 659-694.

Aghion, P. and P. Howitt (1992), "A Model of Growth Through Creative Destruction," Econometrica 60: 323-351.

Aghion, P. and P. Howitt (1998), Endogenous Growth Theory (MIT Press, Cambridge).

Auerbach, A.J. (1996), "Tax Reform, Capital Allocation, Efficiency and Growth," in: H.J. Aaron and W.G. Gale, eds., Economic Effects of Fundamental Tax Reform (Brookings Institution Press, Washington, D.C.) 29-73.

Barro, R.J., and J.W. Lee (2000), "International Data on Educational Attainment: Updates and Implications," National Bureau of Economic Research Working Paper 7911, Cambridge, Massachusetts.

Barro, R.J. and X. Sala-i-Martin (1995), Economic Growth (McGraw-Hill, New York).

Baumol, W. J. (1986), "Productivity Growth, Convergence and Welfare: What the Long-Run Data Show," American Economic Review 76: 1072-1085.

Becker, G.S., T.J. Philipson, and R.R. Soares (2003), "The Quantity and Quality of Life and the Evolution of World Inequality," National Bureau of Economic Research Working Paper 9765, Cambridge, Massachusetts.

Bils, M. and Klenow, P.J. (2000), "Does Schooling Cause Growth?" American Economic Review 90: 1160-1183.

Caves, R.E. (1996), Multinational Enterprise and Economic Analysis, second edition, (Cambridge University Press, New York).

Coe, D.T. and E. Helpman (1995), "International R&D Spillovers," European Economic Review 39: 859-887.

Coe, D.T., E. Helpman, and A.W. Hoffmaister (1997), "North-South R&D Spillovers," Economic Journal 107: 134-139.

De Long, J.B. (1988), "Productivity Growth, Convergence and Welfare: Comment," American Economic Review 78: 1138-1154.

Diamond, J. (1997), Guns, Germs, and Steel: The Fates of Human Societies (WW Norton and Company, New York).

Easterly, W. (2001a), The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics (MIT Press, Cambridge).

Easterly, W. (2001b), "The Lost Decades: Explaining Developing Countries' Stagnation in Spite of Policy Reform 1980-1998," Journal of Economic Growth 6: 135-157.

Easterly, W., M. Kremer, L. Pritchett, and L.H. Summers (1993), "Good Policy or Good Luck? Country Growth Performance and Temporary Shocks," Journal of Monetary Economics 32: 459-484.

Easterly, W. and R. Levine (2001), "It's Not Factor Accumulation: Stylized Facts and Growth Models," World Bank Economic Review 15: 177-219.

Eaton, J. and S. Kortum (1996), "Trade in Ideas: Patenting and Productivity in the OECD," Journal of International Economics 40: 251-278.

Evenson, R.E. and D. Gollin (2003), "Assessing the Impact of the Green Revolution, 1960-2000," Science 300: 758-762.

Fischer, S. (1988), "Symposium on the Slowdown in Productivity Growth," Journal of Economic Perspectives 2: 3-7.

Gollin, D. (2002), "Getting Income Shares Right," Journal of Political Economy 110: 458-474.

Griliches, Z. (1992), "The Search for R&D Spillovers," Scandinavian Journal of Economics 94: S29-S47

Grossman, G.M. and E. Helpman (1991), Innovation and Growth in the Global Economy (MIT Press, Cambridge).

Hall, B.H. (1996), "The Private and Social Returns to Research and Development," in: B.L.R. Smith and C. Barfield, eds., Technology, R&D, and the Economy (Brookings Institution, Washington, D.C.) 140-183.

Hall, B.H. and J. Van Reenen (2000), "How Effective Are Fiscal Incentives for R&D? A Review of the Evidence," Research Policy 29: 449-469.

Hall, R.E. and C.I. Jones (1999), "Why Do Some Countries Produce So Much More Output Per Worker Than Others?" Quarterly Journal of Economics 114: 83-116.

Hendricks, L. (2002), "How Important is Human Capital for Development? Evidence from Immigrant Earnings," American Economic Review 92: 198-219.

Heston, A., R. Summers and B. Aten (2002), Penn World Table Version 6.1, Center for International Comparisons at the University of Pennsylvania (CICUP).

Howitt, P. (1999), "Steady Endogenous Growth with Population and R&D Inputs Growing," Journal of Political Economy 107: 715-730.

Howitt, P. (2000), "Endogenous Growth and Cross-Country Income Differences," American Economic Review 90: 829-846.

Hummels, D.and P.J. Klenow (2004), "The Variety and Quality of a Nation's Exports," forthcoming in the American Economic Review.

Jones, C.I. (1995) "R&D-Based Models of Economic Growth," Journal of Political Economy 103: 759-84.

Jones, C.I. (2001), "Was an Industrial Revolution Inevitable? Economic Growth Over the Very Long Run," Advances in Macroeconomics 1: 1-43.

Jones, C.I. (2002) "Sources of U.S. Economic Growth in a World of Ideas," American Economic Review 92: 220-239.

Jones, C.I. (2004) "Growth and Ideas," Chapter 10 in this volume.

Jones, C.I., and J.C. Williams (1998), "Measuring the Social Return to R&D," Quarterly Journal of Economics 113: 1119-1135.

Jones, L. and R. Manuelli (1990), "A Convex Model of Equilibrium Growth: Theory and Policy Implications," Journal of Political Economy 98: 1008-1038.

Keller, W. (2002), "Geographic Localization of International Technology Diffusion," American Economic Review 92: 120-142.

Keller, W. (2004), "International Technology Diffusion," Journal of Economic Literature 42: 752-782.

King, M.A. and D. Fullerton (1984), The Taxation of Income from Capital: A Comparative Study of the U.S., U.K., Sweden and West Germany (University of Chicago Press, Chicago).

Klenow, P. J. and A. Rodríguez-Clare (1997), "The Neoclassical Revival in Growth Economics: Has It Gone Too Far?" in B. Bernanke and J. Rotemberg, eds., 1997 NBER Macroeconomics Annual (MIT Press, Cambridge) 73-103.

Kortum, S. (1997), "Research, Patenting, and Technological Change," Econometrica 65: 1389-1419.

Kremer, M., A. Onatski, and J. Stock (2001), "Searching for Prosperity," Carnegie-Rochester Conference Series on Public Policy 55: 275-303.

Kremer, M. (1993), "Population Growth and Technological Change: One Million B.C. to 1990," Quarterly Journal of Economics 108: 681-716.

Lederman, D. and W.F. Maloney (2003), "R&D and Development," unpublished paper. Office of the Chief Economist for LCR, The World Bank, Washington, DC.

Lederman, D. and Laura Saenz (2003), "Innovation around the World: A Cross-Country Data Base of Innovation Indicators," unpublished paper, Office of the Chief Economist for LCR, The World Bank, Washington, DC.

Lucas, R.E. (1978), "On the Size Distribution of Business Firms," Bell Journal of Economics: 508-523.

Lucas, R.E. (1988), "On the Mechanics of Economic Development," Journal of Monetary Economics 22: 3-42.

Lucas, R.E. (1990), "Why Doesn't Capital Flow from Rich to Poor Countries?" American Economic Review 80: 92-96.

Lucas, R.E. (2002), Lectures on Economic Growth (Harvard University Press, Cambridge).

Lucas, R.E. (2004), "Life Earnings and Rural-Urban Migration," Journal of Political Economy 112: S29-S59.

Mankiw, N.G., D. Romer, and D.N. Weil (1992), "A Contribution to the Empirics of Economic Growth," Quarterly Journal of Economics 107: 407-437.

Parente, S.L. and E.C. Prescott (1994), "Barriers to Technology Adoption and Development," Journal of Political Economy 102: 298-321.

Parente, S.L. and E.C. Prescott (2004), "A Unified Theory of the Evolution of International Income Levels," Chapter 31 in this volume.

Pritchett, L. (1997), "Divergence, Big Time," Journal of Economic Perspectives 11 (Summer), 3-17.

Pritchett, L. (2004), "Does Learning to Add Up Add Up? The Returns to Schooling in Aggregate Data," forthcoming in the Handbook of Education Economics.

Psacharopoulos, G. and H.A. Patrinos (2002), "Returns to Investment in Education: A Further Update," World Bank Policy Research Working Paper 2881.

Rebelo, S. (1991), "Long Run Policy Analysis and Long Run Growth," Journal of Political Economy 99: 500-521.

Rivera-Batiz, L.A. and P.M. Romer (1991), "Economic Integration and Endogenous Growth," Quarterly Journal of Economics 106: 531-555.

Romer, P.M. (1986), "Increasing Returns and Long-Run Growth," Journal of Political Economy 94: 1002-1037.

Romer, P.M. (1990), "Endogenous Technological Change," Journal of Political Economy 98: S71-S102.

Romer, P.M. (1994), "New Goods, Old Theory, and the Welfare Costs of Trade Restrictions," Journal of Development Economics 43: 5-38.

Solow, R.M. (1956), "A Contribution to the Theory of Economic Growth," Quarterly Journal of Economics 70: 65-94.

Stokey, N.L. (1988), "Learning by Doing and the Introduction of New Goods," Journal of Political Economy 96: 701-717.

Stokey, N.L. (1991), "Human Capital, Product Quality, and Growth," Quarterly Journal of Economics 106: 587-617.

Tamura, R.F. (1991), "Income Convergence in an Endogenous Growth Model," Journal of Political Economy 99: 522-540.

Young, A. (1998), "Growth without Scale Effects," Journal of Political Economy 106: 41-63.

Human Capital and Technology Diffusion*

Jess Benhabib and Mark M. Spiegel†

December 9, 2002

Abstract

This paper generalizes the Nelson-Phelps catch-up model of technology diffusion. We allow for the possibility that the pattern of technology difusion can be exponential, which would predict that nations would exhibit positive catch-up with the leader nation, or logistic, in which a country with a sufficiently small capital stock may exhibit slower total factor productivity growth than the leader nation.

We derive a nonlinear specification for total factor productivity growth that nests these two specifications. We estimate this specification for a cross-section of nations from 1960 through 1995. Our results support the logistic specification, and are robust to a number of sensitivity checks.

Our model also appears to predict slow total factor productivity growth well. 22 of the 27 nations that we identify as lacking the critical human capital levels needed to achieve faster total factor productivity growth than the leader nation in 1960 did achieve lower growth over the next 35 years.

J.E.L. Classification Number: O4

Keywords: human capital, technology diffusion

*Send correspondence to Mark M. Spiegel, Economic Research, Federal Reserve Bank of San Francisco, 101 Market St., San Francisco, CA, 94105, mark.spiegel@sf.frb.org, (415)-974-3241.

†Very helpful comments were received from Richard Dennis, Rody Manuelli, Chris Papageorgiou, and seminar participants at LSU, USC, and the SIEPR/FRBSF Conference on Technical Change. Edmund Chiang provided excellent research assistance. The opinions in this paper are the author's own, and do not necessarily reflect those of the Federal Reserve Bank of San Francisco, or the Board of Governors of the Federal Reserve. We thank the C. V. Starr Center at NYU for technical assistance.

1. Introduction

In a short paper in 1966 Nelson and Phelps offered a new hypothesis to explain economic growth. Their explanation had two distinct components. The first component postulated that while the growth of the technology frontier reflects the rate at which new discoveries are made, the growth of total factor productivity depends on the implementation of these discoveries, and varies positively with the distance between the technology frontier and the level of current productivity. Applied to the diffusion of technology between countries, with the country leading in total factor productivity representing the technology frontier, this is a formalization of the catch-up hypothesis that was originally proposed by Gerschenkron (1962). The second component of the Nelson-Phelps hypothesis suggested that the rate at which the gap between the technology frontier and the current level of productivity is closed depends on the level of human capital. This was a break with the view that human capital is an input into the production process. Nelson and Phelps make this point starkly in the concluding sentence of their paper: "Our view suggests that the usual, straightforward insertion of some index of educational attainment in the production function may constitute a gross mis-specification of the relation between education and the dynamics of production."

The catch-up or technology diffusion component of the Nelson-Phelps hypothesis raises a basic question. If a country, or a firm within an industry, has to incur costs in order to innovate, then why should it not sit back and wait for technology diffusion that flows costlessly? Modern theories of economic growth have paid a great deal of attention to the incentives for innovation and to the market structures that are necessary to sustain R&D. Inventions are typically assumed to give rise to new (often intermediate) products which generate monopoly rents over their lifetime. These rents provide the financial incentives to innovate and to cover the costs of innovation. The costs of invention typically reflect the wages or the patent incomes of researchers. The labor markets allocate workers between

research and production, and in certain cases the allocation of workers across different occupations can involve decisions to acquire costly human capital. When a vintage structure is present, newer and technologically more efficient intermediate goods or production processes may coexist with older ones that remain inside the technology frontier. A critical by-product of an innovation, not captured by the monopoly rents that it generates, is the expansion of the stock of basic knowledge This basic knowledge, freely available to all, enhances the productivity of future research, facilitates future innovations and is the source of scale effects.

In the Nelson-Phelps framework, disembodied technical know-how flows from the technology leader to its followers and augments their total factor productivity. Patent protection or blueprint ownership is not explicitly postulated, and therefore an alternative mechanism must be in operation to sustain inventive activity and to prevent free-riding. A number of models have directly addressed the impact of imitation that dissipates rents on innovative activity by explicitly introducing costs of imitation. In an early investigation by Grossman and Helpman (1991, Chapter 11, see also Helpman (1993), Segerstrom (1991)), the North, where patent protection is in effect, innovates, and the South, where labor costs are lower, imitates at a cost. Aghion, Harris and Vickers (1997), building on Grossman and Helpman (1991), suggest a leapfrogging model where firms can, by incurring an appropriate cost, catch-up and overtake their rivals to capture a larger share of the profits. Eaton and Kortum (1999) construct a model with patenting costs where patens decrease but not eliminate the hazard of imitation. To construct an equilibrium with technology diffusion, Barro and Sala-i Martin [1995, also (1997)] introduce a model where in the leading country the costs of innovation are low relative to the costs of imitation, while in the follower country the reverse is true. Basu and Weil (1998) propose a model where technological barriers to imitation in the South arise from significant differences in factor proportions between North and South, with the possible emergence of "convergence clubs." Such differences in

endowments may not provide the most "appropriate" opportunities for imitation, and fail to direct technical change towards efficient cost savings (see Acemoglu (2002)). Technology may nevertheless flow between convergence clubs, with imitation costs rather than patent protection sustaining innovative activity within the clubs. Eeckhout and Jovanovic (2002) construct a model where imitators can implement technology only with a lag, and this implicit imitation cost means that innovators find it optimal to maintain their lead. It seems clear then that some costs of imitation and certain advantages to innovation must be present if technology diffusion is to play a role in economic growth. Therefore, underlying the Nelson-Phelps model there must be an appropriate market structure with an economic equilibrium that sustains innovative activity in the face of technology diffusion.

The empirical literature on technology diffusion has been growing, despite difficulties in measurements. The survey of Griliches (1992) lends support to the view that there are significant R&D spillovers. Coe and Helpman (1995) find that R&D abroad benefits domestic productivity, possibly through the transfer of technological know-how via trade. Branstetter (1996), looking at disaggregated data, finds research spillovers across firms that are close in "technology space." Nadiri and Kim (1996) suggest that the importance of research spillovers across countries varies with the country: domestic research seems important in explaining productivity in the US but the contribution of foreign research is more important for countries like Italy or Canada. The role of human capital in facilitating technology adoption is documented by Welch (1975), Bartel and Lichtenberg (1987) and Foster and Rosenzweig (1995). Benhabib and Spiegel (1994), using cross-country data, investigate the Nelson-Phelps hypothesis and conclude that technology spills over from leaders to followers, and that the rate of the flow depends on levels of education. In fact a good deal of the recent empirical literature has focused on whether the level of education speeds technology diffusion and leads to growth, as

suggested by Nelson Phelps, or whether education acts as a factor of production, either directly or through facilitating technology use. (See for example, Islam (1995), Eaton and Kortum (1996), Temple (1999), Krueger and Lindahl (2001), Pritchett, Klenow and Rodriguez-Clare (1997), Hall and Jones (1999), Bils and Klenow (2000), Duffy and Papageorgiou (2000), and Hanushek and Kimko (2000)).

The policy implications of distinguishing between the role of education as a factor of production and a factor that facilitates technology diffusion are significant. In the former, the benefit of an increase in education is its marginal product. In the latter, because the level of education affects the growth rate of total factor productivity and technology diffusion, its benefit will be measured in terms of the sum of its impact on all output levels in the future. Following Nelson and Phelps (1966), in Benhabib and Spiegel (1994) we characterize the latter relationship through a specification to explain growth that includes a term interacting the stock of human capital with backwardness, measured as a country's distance from the technology leader.

There are potentially important implications of distinguishing between different functional forms for the technology diffusion process. The technology diffusion process specified by Nelson and Phelps and widely used in the literature is known as the confined exponential diffusion [Banks (1994)]. An alternative diffusion process is the logistic model of technology diffusion. A priori, there appears to be no reason to favor one of these technology diffusion specifications over the other, and their specification appears to differ very little. Nevertheless, as we demonstrate below, these specifications can have very different implications for a nation's growth path: For the exponential diffusion process, the steady state is, for all parametrizations, a balanced growth path, with all followers growing at the pace determined by the leader nation that acts as the locomotive. In contrast, the logistic model allows for a dampening of the diffusion process so that the gap between the leader and a follower can keep growing. Indeed, we demonstrate that

if the human capital stock of a follower is sufficiently low, the logistic diffusion model implies divergence in total factor productivity growth rates, not catch-up. On this point, also see Howitt and Mayer-Foulkes (2002).

Below we derive an empirical specification that nests these two forms of technology diffusion in a model where total factor productivity growth depends on initial backwardness relative to the stock of potential world knowledge, proxied in our model as the total factor productivity level of the leader country. We then test this specification for a cross-section of total factor productivity growth of 84 countries from 1960 through 1995. We obtain robust results supporting a positive role for human capital as an engine of innovation, as well as a facilitator of catch-up in total factor productivity.

As our results favor a logistic form of technology diffusion, some countries may indeed experience divergence in total factor productivity growth. To investigate this result, we derive a point estimate from our estimation results for the minimum initial human capital level necessary to exhibit catch-up in total factor productivity relative to the leader nation, which is the United States in our sample. The point estimate in our favored specification indicates that an average of 1.78 years of schooling was required in 1960 to achieve convergence in total factor productivity growth with the United States.

Under this criterion, we identify 27 countries in our sample that our point estimates predict will exhibit slower total factor productivity growth than the United States. Our data shows that over the next 35 years, 22 of these 27 countries did indeed fall farther behind the United States in total factor productivity, while the remaining bulk of the nations in our sample exhibited positive catch-up in total factor productivity. While this result is not a formal test of our model, its ability to correctly identify countries that would subsequently exhibit slower total factor productivity growth than the United States is reassuring.

We then repeat our exercise using 1995 figures to identify the set of nations

that are still falling behind in total factor productivity growth. Because the United States had higher education levels in 1995, we estimate a higher threshold level for total factor productivity growth convergence with the United States. Our estimate was that 1.95 average years of schooling in the population over the age of 25 was necessary for faster total factor productivity growth than the leader nation. Fortunately, the higher overall education levels achieved by most countries over the past 35 years left few countries falling the threshold levels in education to achieve catch-up in growth rates. We identified only four countries as still below the threshold in 1995: Mali, Mozambique, Nepal, and Niger. With the exception of these four nations, our results indicate that most of the world is not in a permanent development trap, at least in terms of total factor productivity growth. Nevertheless, it should be pointed out that catch-up in total factor productivity is not a guarantee of convergence in per capita income, as nations must also be successful in attracting physical capital to achieve the latter goal.

The remainder of the paper is divided into five sections. Section 2 introduces the exponential and logistic specifications of the Nelson-Phelps model and examines their steady-state implications. Section 3 compares the diffusion models with that of Barro and Sala-i-Martin (1997). Section 4 derives a non-linear growth specification that nests the exponential and logistic technology diffusion functional forms. Section 5 estimates this model using maximum likelihood for a cross-section of countries. Section 6 uses the point estimates from our estimation to identify nations that are predicted to fail to exhibit divergence in total factor productivity growth in 1960 and 1995. Lastly, Section 7 concludes.

2. Variations on the Nelson-Phelps Model

We will examine the implications of two types processes often studied in the context of disaggregated models of technology diffusion (Banks (1994)). We can

express the original Nelson-Phelps model of technology diffusion as follows:

$$\frac{\dot{A}_i(t)}{A_i(t)} = g\left(H_i\left(t\right)\right) + c\left(H_i\left(t\right)\right)\left(\frac{A_m\left(t\right)}{A_i\left(t\right)} - 1\right) \tag{2.1}$$

where $A_i(t)$ is the TFP, $g_i\left(H_i\left(t\right)\right)$ is the component of TFP growth that depends on the level of education $H_i\left(t\right)$ in country i and $c\left(H_i\left(t\right)\right)\left(\frac{A_m(t)}{A_i(t)} - 1\right)$ represents the rate of technology diffusion from the leader country m to country i. We assume that $c_i\left(\cdot\right)$ and $g_i\left(\cdot\right)$ are increasing functions. The level of education $H_i\left(t\right)$ affects the rate at which the technology gap $\left(\frac{A_m(t)}{A_i(t)} - 1\right)$ is closed. If the ranking of $g_i\left(H_i\left(t\right)\right)$ across countries do not change, or if $H_i's$ are constant, a technology leader will emerge in finite time with $g_m = g\left(H_m\left(t\right)\right) > g\left(H_i\left(t\right)\right) = g_i$. After that the leader will grow at rate g_m and the followers will fall behind in levels of TFP until the point at which their growth rate will match the leader's growth rate g_m This can be seen from the solution of the above equation when H_i's are constant [1]:

$$A_i\left(t\right) = \left(A_i\left(0\right) - \Omega A_m\left(0\right)\right)e^{(g_i - c_i)t} + \Omega A_m\left(0\right)e^{g_m t} \tag{2.2}$$

where $c_i = c\left(H_i\right)$, $g_i = g\left(H_i\right)$ and

$$\Omega = \frac{c_i}{c_i - g_i + g_m} > 0.$$

It is clear, since $g_m > g_i$, that

$$\lim_{t\to\infty} \frac{A_i\left(t\right)}{A_m\left(t\right)} = \Omega$$

[1] The general solution when H_i's are not constant is given by:

$$A_i\left(t\right) = A_i\left(0\right)e^{-\int_0^t (g(H_i(s)) - c(H_i(s)))ds}$$
$$\cdot\left[1 + \frac{1}{A_i\left(0\right)}\left(\int_0^t c\left(H_i\left(\tau\right)\right)\left(A_m\left(0\right)e^{\int_0^\tau g(H_m(\zeta))d\zeta}\right)e^{\int_0^\tau (g(H_i(\xi)) - c(H_i(\xi)))d\xi}d\tau\right)\right]$$

This is, for all parametrizations, a world balanced growth path with the leader acting as the "locomotive." Technology diffusion and "catch-up" assures that despite scale effects and educational differences, all countries eventually grow at the same rate.[2]

The technology diffusion and catch-up processes outlined above are also known as the confined exponential diffusion process (see Banks(1994)) An alternative formulation that is similar in spirit is the logistic model of technology diffusion (see Sharif and Ramanthran (1981)). It is given by

$$\begin{aligned}\frac{\dot{A}_i(t)}{A_i(t)} &= g\left(H_i\left(t\right)\right) + c\left(H_i\left(t\right)\right)\left(1 - \frac{A_i\left(t\right)}{A_m\left(t\right)}\right) \\ &= g\left(H_i\left(t\right)\right) + c\left(H_i\left(t\right)\right)\left(\frac{A_i\left(t\right)}{A_m\left(t\right)}\right)\left(\frac{A_m\left(t\right)}{A_i\left(t\right)} - 1\right)\end{aligned} \tag{2.3}$$

The difference of the dynamics under the logistic model of technology diffusion and the confined exponential one is due to the presence of the extra term $\left(\frac{A_i(t)}{A_m(t)}\right)$. This term acts to dampen the rate rate of diffusion as the distance to the leader increases, reflecting perhaps the difficulty of adopting distant technologies. As shown by Basu and Weil (1998), the frontier technology may not be immediately "appropriate" for the follower if differences in factor proportions between leader and follower are large. We may observe convergence clubs, as documented by Durlauf and Johnson (1995), from which follower countries can break out only by investing in physical and human capital. Catch-up therefore may be slower when the leader is either too distant or too close, and is fastest at intermediate distances.[3]

If we assume, as before, that H_i's (and therefore, c_i's and g_i's) are constant

[2] Note however that in transition, the higher is initial $A_i\left(0\right)$, the smaller is the technology gap to the leader and therefore the slower is the growth. This negative dependence on initial conditions is similar to standard convergence results in the neoclassical growth model, but the logic of catch-up is different.

[3] An alternative view of technology adoption through diffusion that follows a logistic pattern

such that $H_m > H_i$, and therefore that $c(H_m) > c(H_i)$, then the solution to the logistic technology diffusion equation is given by [4][5][6]

$$A_i(t) = \frac{A_i(0)\, e^{(g_i+c_i)t}}{\left(1 + \frac{A_i(0)}{A_m(0)} \frac{c_i}{(c_i+g_i-g_m)} \left(e^{(c_i+g_i-g_m)t} - 1\right)\right)} > 0 \tag{2.4}$$

This equation can be written as

$$A_i(t) = \frac{A_m(0)\, e^{g_m t}}{\left(e^{-(c_i+g_i-g_m)t} \left(\frac{A_m(0)}{A_i(0)} - \frac{c_i}{(c_i+g_i-g_m)}\right) + \frac{c_i}{(c_i+g_i-g_m)}\right)} \tag{2.5}$$

so that in the limit,

$$\lim_{t\to\infty} \frac{A_i(t)}{A_m(t)} = \left\{ \begin{array}{ccc} \frac{(c_i+g_i-g_m)}{c_i} & & (c_i+g_i-g_m) > 0 \\ \frac{A_i(0)}{A_m(0)} & if & (c_i+g_i-g_m) = 0 \\ 0 & & (c_i+g_i-g_m) < 0 \end{array} \right\}. \tag{2.6}$$

Equation (2.6) implies that in the case of the logistic diffusion model, the steady state growth relationship will depend on the relative magnitude of the

borrows from epidemiology. The rate of adoption in a fixed population may depend on the rate of contact between adopters and hold-outs (those that are infected and those that are healthy). The adoption rate is highest when there are an equal number of both types, and lower when there is either a small or a large proportion of adopters. Also observing the successes and implementation errors of the first adopters, together with the competitive pressures that first adopters create, may result in a speeding up of adoption rates. See Mansfield (1968).

[4] Provided that $(c_i + g_i - g_m) \neq 0$. If $(c_i + g_i - g_m) = 0$, then the equation reduces to exponential form $A_i(t) = A_i(0)e^{(g_i+c_i)t}$.

[5] The general solution where H_i's are functions of time can be computed by defining $B_i = (A_i)^{-1}$ and transforming the logistic form into the confined exponential. After some computations, the general form can be obtained as

$$A_i(t) = \frac{A_i(0)\, e^{\int_0^t (g(H_i(s))+c(H_i(s)))ds}}{\left(1 + A_i(0)\left(\int_0^t c(H_i(\tau))\left(\left(A_m(0)^{-1}\right) e^{-\int_0^\tau g(H_m(\zeta))d\zeta}\right) e^{\int_0^\tau (g(H_i(\xi))+c(H_i(\xi)))d\xi} d\tau\right)\right)}$$

[6] $A_i(t) > 0$ because when $c_i + g_i - g_m \neq 0$, $\frac{c_i}{(c_i+g_i-g_m)}\left(e^{(c_i+g_i-g_m)t} - 1\right) > 0$.

catch-up rate and the difference in the growth rate due to innovation, $g_m - g_i$. If the catch-up rate exceeds the differential growth rate solely due to educational differences between the leader and follower, that is if $c(H_i) + g(H_i) - g(H_m) > 0$, then the leader will have a locomotive effect and pull the followers along. In such a case growth rates will converge. However, if the education level of a follower is so low that $c(H_i) + g(H_i) - g(H_m) < 0$, then the follower will not be able to keep up, growth rates will diverge, and the income ratio of the follower to the leader will go to zero.

This highlights the critical role of the type of technology diffusion process and its interaction with education in fostering economic growth: a country with a low level of education may still keep within the gravitational pull of the technology leader, provided that the level of education is high enough to permit sufficient diffusion. If technology diffusion is of the logistic type, countries with educational levels that are too low will get left behind and we may observe the phenomenon of "convergence clubs." Escaping from the lower "club" is nevertheless possible through investments in human capital, as discussed by Basu and Weil (1998) [7]. The implications of logistic versus exponential technology diffusion for economic growth can therefore be quite divergent.

Note that we can append the Nelson-Phelps framework, either in the logistic or the confined exponential form, to the Romer (1990) model by adding the catch-up term the research sector producing the blueprints A. The marginal product of H in the research sector will now reflect an effect from the catch-up term, and increase the allocation of H towards the research sector away from production or leisure. If, as in Romer, we assume that H is constant while knowledge, A, is accumulated, and also assume that goods use labor but not H, we may focus on the allocation of H to imitation through catch-up or to innovation. Adopting a

[7] We should note that $c(H_i)$ may also depend on barrers to innovation as in Parente and Prescott (1994), so that in fact we have $c(H_i, X)$, where X represents the level of barriers.

linear specification with $g(H_i) = gH_i$, $c(H_i) = cH_i$, the marginal product of H in innovation is given by $gA_i(t)$ while in imitation, for the confined exponential case, it is $cA_i(t)\left(\frac{A_m(t)}{A_i(t)} - 1\right)$. These marginal products are independent of H_i so we may have a bang-bang solution, with all of H_i allocated towards catch-up and imitation up to a threshold, and to innovation otherwise[8]. In what follows we will, for he time being, abstract from issues regarding the allocation of H_i, and assume that all of it enters both imitation and catch-up as a non-excludable public good.

3. Some Microfoundations based on the diffusion model of Barro and Sala-i-Martin

To set the stage first we express the confined exponential and logistic growth equations discussed above in stationary variables by defining

$$B(t) = \frac{A_i(t)}{A^*(0)} e^{-g_m t} \tag{3.1}$$

for all i. Then, for the logistic case, we have

$$\begin{aligned} \frac{\dot{B}}{B} &= c(H_i)(1 - B) + g(H_i) - g(H_m) \\ \dot{B} &= (c(H_i) + g(H_i) - g(H_m))B - c(H_i)B^2 \end{aligned} \tag{3.2}$$

If $H_i's$ are fixed the solution is,

$$B(t) = \left(\frac{c_i + g_i - g_m}{c_i}\right)\left[1 + \left(\left(\frac{c_i + g_i - g_m}{c_i}\right)\left(\frac{A^*(0)}{A(0)}\right) - 1\right)e^{-\left(\frac{c_i + g_i - g_m}{c_i}\right)t}\right]^{-1} \tag{3.3}$$

[8] A further consideration is the allocation of resources between imitative and innovatiove uses, where the efficient allocation changes as the distance to the technology frontier narrows. The market allocation may differ from the efficient allocation due to to a variety of factors, and policy interventions may improve welfare. See Acemoglu, Aghion and Zilibotti (2002).

So if $c_i + g_i - g_m > 0$,

$$\lim_{t\to\infty} B(t) = \left(\frac{c_i + g_i - g_m}{c_i}\right),$$

while if $c_i + g_i - g_m < 0$, $\lim_{t\to\infty} B(t) = 0$.[9] Note from equation (3.2) that in the latter case where $c_i + g_i - g_m < 0$, there is no steady state with $B > 0$.

In the confined exponential case

$$\begin{aligned}\frac{\dot{B}}{B} &= c(H_i)\left(B^{-1} - 1\right) + g(H_i) - g(H_m)\\ \dot{B} &= c(H_i) - \left(c(H_i) + g(H_m) - g(H_i)\right)B \qquad (3.4)\end{aligned}$$

Since $c(H_i) + g(H_m) - g(H_i) > 0$, it is clear from (3.4) that there exists a stable steady state at $B = \frac{c(H_i)}{c(H_i)+g(H_m)-g(H_i)}$.

In the Barro and Sala-i-Martin (1997) model, the North, where innovation is cheap, is the leader. It innovates by introducing new intermediate goods, and receives no diffusion through imitation from the South. As in a typical growth model of the Romer type, it grows at a constant rate γ.The South introduces new intermediate goods through imitation. In both countries the production of final goods is given by:

$$Y_i = A_i (L_i)^{1-\alpha} \sum_{j=1}^{N_i} (X_{ij})^{\alpha} \qquad i = 1, 2$$

where the North is country 1 and the South is country 2, so that $N_1 > N_2$. The profits of the $j'th$ intermediate goods producer is given by $\pi_{2j} = (P_{2j} - 1) X_{2j}$ where P_{2j} is the price of the intermediate good in terms of the final good in the South. The cost of imitation in the South is $v_2\left(\frac{N_2}{N_1}\right)$.In a symmetric equilibrium investment in R&D is given by

$$v_2 \dot{N}_2 = Y_2 - C_2 - N_2 X_2 \qquad (3.5)$$

[9] In the case $c_i + g_i - g_m > 0$, $B(t)$ should (if the assumption that $H_i's$ are constant holds) exhibit the S-shaped logistic diffusion.

where the LHS is the cost of introducing a new intermediate good through imitation, and the RHS is income minus consumption minus the cost of operating the existing intermediate goods (since $X_{i2} = X_2$ for all i). Barro and Sala-i-Martin show that in equilibrium X_2 and $\frac{Y_2}{N_2}$ are constants.[10], [11] For simplicity of exposition we will also assume a constant consumption propensity out of income, so that $C_2 = \mu (Y_2 - NX_2)$, so that

$$\frac{\dot{N}_2}{N_2} = \frac{1}{v_2}\left(\frac{Y_2}{N_2} - X_2\right)(1-\mu) \equiv \frac{1}{v_2}P$$

and

$$\frac{\dot{B}}{B} = \frac{1}{v_2}P - \gamma \tag{3.6}$$

where $B = \frac{N_2}{N_1}$.[12].

Barro and Sala-i-Martin assume that

$$v_2 = \eta\left(\frac{N_2}{N_1}\right)^{\sigma} \equiv \eta B^{\sigma}, \qquad \sigma > 0$$

Imitations costs are higher, the closer the follower is to the leader. We can now assume that η depends negatively on human capital, so that the cost of imitation declines with H. Introducing this specification into (3.6) we get

$$\dot{B} = \eta^{-1}B^{1-\sigma}P - \gamma B = B\left(\eta^{-1}PB^{-\sigma} - \gamma\right)$$

[10] In particular, $X_2 = L_2 (A_2)^{\frac{1}{1-\alpha}} (\alpha)^{\frac{2}{1-\alpha}}$ and $\frac{Y_2}{N_2} = (A_2)^{\frac{1}{1-\alpha}} \alpha^{\frac{2\alpha}{1-\alpha}} L_2$ where L_2 is, for simplicity, the constant the labor supply in the South.

[11] In BSM, consumption growth depends on the interest rate, which reflects the value of the stream of profits divided by the cost of imitation. Since the cost of imitation depends on N_2/N_1, the dynamic system is two-dimensional in N_2/N_1 and C_2/N_1. For details, see Barro and Sala-i-Martin (1997).

[12] To see that $Y_2 - N_2X_2$ corresponds to income note that in a symmetric equilibrium $\left(Y_2 - \int_0^{N_2} P_{2j}X_{2j}dj\right) + \int_0^{N_2} \pi_{2j}dj = Y_2 - N_2(P_2X_2 - \pi_2)$. Thus we must show that $P_2X_2 - \pi_2 = X_2$ where π_2 is profits. Since in equilibrium $\pi_2 = \left(\alpha^{-1} - 1\right)\left(\alpha^2 AL_2^{1-\alpha}\right)^{\frac{1}{1-\alpha}}$, $X_2 = L_2 (A_2)^{\frac{1}{1-\alpha}} (\alpha)^{\frac{2}{1-\alpha}}$ and $P_2 = \alpha^{-1}$, we have $P_2X_2 - \pi_2 = \alpha^{-1}\left(\alpha^2 AL_2^{1-\alpha}\right)^{\frac{1}{1-\alpha}} - \left(\alpha^{-1} - 1\right)\left(\alpha^2 AL_2^{1-\alpha}\right)^{\frac{1}{1-\alpha}} = \left(\alpha^2 AL_2^{1-\alpha}\right)^{\frac{1}{1-\alpha}} = X_2$

which has a stable steady state at $B = \left(\frac{\eta\gamma}{P}\right)^{-\frac{1}{\sigma}}$. Therefore this specification of imitation costs yields the same qualitative conclusions as the confined exponential diffusion used by Nelson and Phelps: the leader acts as the engine of growth pulling the followers along.

We now modify the imitation costs to correspond to the case of logistic technology diffusion. Let

$$v_2 = \eta (1 - B)^{-1}$$

where again v_2 is increasing in B. Now the diffusion equation becomes

$$\dot{B} = \eta^{-1} P (1 - B) B - \gamma B = (\eta^{-1} P - \gamma) B - \eta^{-1} P B^2 \tag{3.7}$$

which has the same logistic structure as (3.2). In particular, there is a positive steady state $B = \frac{\eta^{-1}P - \gamma}{\eta^{-1}P}$ only if $\eta^{-1}P > \gamma$ Otherwise B converges to 0, and there is no catch-up. More generally since the non-zero steady state is given by $v_2(B^*) = \frac{P}{\gamma}$, if $v_2(0) > \frac{P}{\gamma}$ there will not be a positive steady state $B^* > 0$, but the steady state $B = 0$ will be stable.[13,14]

If η is decreasing in H so that imitation costs decline with human capital, for sufficiently low levels of H we may have $\eta^{-1}P < \gamma$, and the South will never catch

[13] If on the other hand, we adopt the confined exponetial $\left(\nu_2(B) = \eta\left(B^{-1} - 1\right)\right)$ or the Barro and Sala-i Martin specification $(\nu_2(B) = \eta B^{\sigma})$, then $\nu_2(0) = 0$, so that the diffusion rate approaches infinity and the imitation costs go to zero when N_2 and B tend to zero, a strong and unlikely assumption.

[14] In Barro and Sala-i Martin's more general model consumption growth, given by the Euler equation, depends on the rate of return on intermediate goods, which varies through time with the distance to the frontier. At a steady state with $B > 0$, we obtain $\frac{C_2}{N_2} = \frac{(1+\alpha)}{\alpha}\pi_2 - \gamma v_2(B)$ where π_2 is the profit rate. Plugging this into (3.5) and simplifying, we get $v_2(B) = \frac{\pi_2}{(\theta\gamma+\rho)}$ where ρ is the discount rate and θ^{-1} is the intertemporal consumption elasticity. For the Barro and Sala-i Martin specification, $v_2(B) = \eta B^{\sigma},\ v_2(0) = 0$. If however $v_2(0) > \frac{\pi_2}{(\theta\gamma+\rho)}$, no positive steady state B exists. This is likely if v_2 also depends (inversely) on human capital and if human capital levels are low.

up in growth rates. In such circumstances there may be incentives to accumulate human capital. If however there are market imperfections in the accumulation of capital, or if H mostly provides external effects, there may not exist sufficient market incentives for the accumulation of H, so that subsidies to education may be necessary to improve growth and welfare.

4. A nested specification

We can also, for purposes of estimation, specify a diffusion process that nests the logistic and confined exponential diffusion processes. Using the definition of B given in (3.1), we can modify (3.2) as

$$\frac{\dot{B}}{B} = \frac{c(H_i)}{s}\left(1 - B^s\right) + g(H_i) - g(H_m) \tag{4.1}$$

$$\dot{B} = \left(\frac{c(H_i) + sg(H_i) - sg(H_m)}{s}\right) B - \frac{c(H_i)}{s} B^{s+1} \tag{4.2}$$

$$\dot{B} = \left(\frac{c(H_i) + sg(H_i) - sg(H_m)}{s}\right) B \left(1 - \left(\frac{B^s}{\left(1 + \frac{s(g_i - g_m)}{c_i}\right)}\right)\right) \tag{4.3}$$

with $s \in [-1, 1]$. Note that if $s = 1$, this specification collapses to the logistic, and if $s = -1$, it collapses to the confined exponential[15]. In its general form this is a Bernoulli equation, whose solution, when H_i and H_m are constants so that $c_i = c(H_i)$, $g_m = g(H_m)$, $g_i = g(H_i)$, is given by :

$$B(t) = \left(\frac{\left(1 + \frac{s(g_i - g_m)}{c_i}\right)}{\left(1 + \left(\left(1 + \frac{s(g_i - g_m)}{c_i}\right) B(0)^{-s}\right) - 1\right) e^{-(c_i + s(g_i - g_m))t}}\right)^{\frac{1}{s}} \tag{4.4}$$

[15] See Richards (1959).

Since the leader has more human capital, $H_m > H_i$, we have $g_m > g_i$. It follows that if either $c_i + s(g_i - g_m) > 0$,or if $s < 0$,

$$\lim_{t\to\infty} B(t) = \left(1 + \frac{s(g_i - g_m)}{c_i}\right)^{\frac{1}{s}},$$

while if $\left(1 + \frac{s(g_i-g_m)}{c_i}\right) < 0$, and $s > 0$, $\lim_{t\to\infty} B(t) = \lim_{t\to\infty} \frac{A_i(t)}{A_m(t)} = 0$ [16] In the latter case, as noted in the previous section, the South never catches up and growth rates diverge. [17]

To test this nested specification empirically we can specify it as:

$$\Delta a_{it} = \left(g + \frac{c}{s}\right) h_{it} - \frac{c}{s} h_{it} \left(\frac{A_{it}}{A_{mt}}\right)^{s}.$$

where Δa_{it} is the growth of TFP for country i, h_{it} is its initial or average human capital and $\left(\frac{A_{it}}{A_{mt}}\right)$is the ratio of the country′s TFP to that of the leader. Note

[16]If c_i and g_i vary with time because H_i changes with time, (4.1) is the classic Bernoulli equation which we can write as:

$$\dot{B} = f(t) B + g(t) B^{s+1}$$

where $f(t) = \frac{c(H_i(t))}{s} + g(H_i(t)) - g(H_m(t))$and $g(t) = -\frac{c(H_i(t))}{s}$ as in equation (4.2). The solution is:

$$B(t) = \left(Ce^{\phi(t)} + se^{\phi(t)} \int e^{\phi(\tau)} g(\tau)\, d\tau\right)^{-\frac{1}{s}}$$

where $\phi(t) = s\int f(\tau)\, d\tau$ and C is an integration constant such that $C^{-\frac{1}{s}} = B(0)$.

[17]When $s \to 0$, the diffusion process converges to the Gompertz growth model:

$$B = \lim_{s\to 0} \left(1 + \frac{s(g_i - g_m)}{c_i}\right)^{\frac{1}{s}} \exp\left(-e^{k-c_i t}\right) = \exp\left(\frac{(g_i - g_m)}{c_i}\right) \exp\left(-e^{k-c_i t}\right) \tag{4.5}$$

$$\dot{B} = c_i B\, \ell n \left(\frac{\exp\left(\frac{(g_i-g_m)}{c_i}\right)}{B}\right) \tag{4.6}$$

where $e^k = \left(\frac{(g_i-g_m)}{c_i}\right) - \ell n(B(0))$. So $\lim_{t\to\infty} B = \exp\left(\frac{(g_i-g_m)}{c_i}\right) > 0$. To see this note that, using LHopital's Rule, the right side of equation (4.3) collapses to $c_i B\, \ell n \left(\frac{\exp\left(\frac{(g_i-g_m)}{c_i}\right)}{B}\right)$.

again that this specification nests the logistic ($s = 1$) and exponential ($s = -1$) models. As discussed above, the values of c, g and s will determine whether a country will converge to the growth rate of the leader or whether the the growth rates will diverge. In particular, or our linear specification $c(h_{it}) = c_i = ch_{it}$, $g(h_{it}) = g_i = gh_{it}$ and $g(h_{mt}) = g_m = gh_{mt}$, "the catch-up condition" for the growth rate of a country to converge to the growth rate of the leader becomes (for $s \in (0, 1]$) :

$$c^* = 1 + \frac{c}{sg} > \frac{h_{mt}}{h_{it}} \tag{4.7}$$

Countries for which $\left(\frac{h_{mt}}{h_{it}}\right) > c^*$ will not converge to the leader's growth rate unless they invest in their human capital to reverse this inequality.[18]

5. Empirical Evidence

5.1. Measurement of Total Factor Productivity

Data for real income and population growth were obtained from the Penn World Tables, version 6.1.Data for human capital, which is proxied by average years of schooling in the population above 25 years of age, was obtained from the updated version of the Barro Lee (1993) data set. Our sample consists of 85 countries with data for the period 1960-1995. We estimate this sample both as a cross-section of 35 years of growth and as a panel of five-year growth rates.

Physical capital stocks were calculated according to the method used in Klenow and Rodriguez-Clare (1997). Initial capital stocks are calculated according to the

[18] As noted earlier however te catch-up coefficient $c(h_{it})$ may depend on other institutional factors in addition to human capital, like barriers to innovation as in Parente and Prescott (1994). In such a case we may want to modify the catch-up coefficient to $c_i = \alpha_i ch_{it}$ where α_i reflects country speific barriers to innovation.

following formula

$$\frac{K}{Y}_{1960} = \frac{I/Y}{\gamma + \delta + n} \tag{5.1}$$

where I/Y is the average share of physical investment in output from 1960 through 2000, γ represents the average rate of growth of output per capita over that period, n represents the average rate of population growth over that period, and δ represents the rate of depreciation, which is set equal to 0.03. Given initial capital stock estimates, the capital stock of country i in period t satisfies

$$K_{it} = \sum_{j=0}^{t} (1-\delta)^{t-j} I_{ij} + (1-\delta)^{t} K_{1960}. \tag{5.2}$$

Total factor productivity growth was estimated from a constant returns to scale Cobb-Douglas production function with the capital share set at 1/3 and the labor share set at 2/3.[19] For country i in period t

$$a_{it} = y_{it} - \frac{1}{3}k_{it} - \frac{2}{3}l_{it} \tag{5.3}$$

where a_{it} represents the long of total factor productivity, y_{it} represents the log of real output, k_{it} represents the log of the physical capital stock, and l_{it} represents the log of the population.

Total factor productivity estimates for 1960 and 1995, as well as estimates of average annual growth in total factor productivity over the period are shown in Table 1. The results seem pretty intuitive, as the Asian Tiger countries, including Taiwan, Singapore, Korea, Hong Kong and Thailand, lie notably at or near the top in terms of total factor productivity growth, while the five countries exhibiting the lowest growth in total factor productivity are Mozambique, Niger, Central African

[19] Gollin (2002) estimates that the share of labor lies between 0.65 and 0.80 for a cross-section of world economies. Keller (2002) estimated TFP with both the factor shares used above and the capital and labor shares set equal to one-half and obtained similar ordinal rankings of total factor productivity levels across countries.

Republic, Nicaragua, and Zambia. All of these countries experienced negative total factor productivity growth over the sample period, as did Mali, Senegal, Venezuela, Togo and Cameroon. Out of this group of ten negative total factor productivity growth countries, only Venezuela's appearance is surprising, and that can probably be attributed to its buildup of physical capital for oil production. In the case of the five highest total factor productivity growth countries, our results would no doubt differ slightly if our sample included the Asia crisis of 1997. Nevertheless, the set of countries exhibiting high total factor productivity growth seems intuitive as well.

A simple scatter plot of initial human capital levels and subsequent total factor productivity growth over the estimation period is shown in Figure 1. The raw correlation between these two variables is clearly positive, suggesting that nations with larger initial human capital stocks tend to exhibit higher total factor productivity growth holding all else constant. There are a number of interesting outliers. The Asian tiger nations are noteworthy as nations that exhibited fast total factor productivity growth and began the estimation period with relatively stocks of initial human capital.[20] On the other hand, there are a number of countries that exhibited total factor productivity declines that began the period with exceptionally low levels of initial human capital, including Mali, Niger, Togo, Mozambique, and the Central African Republic.

[20]It is unfortunate that our sample ends in 1995, because the Asian "tiger" nations suffered large declines in the 1997 crisis. However, we confirmed that total factor productivity growth of these nations was still exceptionally high for the Asian tiger nations for which longer 39 year data from 1960 to 1999 was available. This included all of the tigers except Singapore.

5.2. Model Specification

As discussed above, the following non-linear cross-sectional specification nests the exponential and logistic functional forms of technology diffusion

$$\Delta a_i = b + \left(g + \frac{c}{s}\right) h_i - \left(\frac{c}{s}\right) h_i \left(\frac{A_i}{A_m}\right)^s + \varepsilon_i \tag{5.4}$$

where Δa_i represents the average annual growth rate in TFP of country i, h_i represents the log of country $i's$ stock of human capital, A_i represents the level of country $i's$ stock of TFP, A_m represents the level of TFP of the leader nation, and ε_i is an i.i.d. disturbance term. The coefficients to be estimated represent b, $\left(g + \frac{c}{s}\right)$, -$\left(\frac{c}{s}\right)$, and s respectively.

We are agnostic as to whether it is appropriate to include the constant term b. This term could be interpreted as exogenous technological progress that is independent of human capital and technology diffusion. It is difficult to envision any type of technological progress that would be common across our sample and completely independent of the levels of national human capital. In the case where "accidental technological progress" truly does take place, it is far more likely that it would appear in our error term as it would be confined to specific nations within our sample. Nevertheless, we report our estimation results both without and with the constant terms included as a measure of their robustness.

Our model nests two alternative hypotheses. First, we have our Nelson-Phelps type model of technology diffusion, dependent on human capital and technological backwardness, that is of the confined exponential type. As we noted above, this model would correspond to the above specification with s equal -1. Second, we have our logistic specification for the technology diffusion process, which would correspond to s being equal to 1. We therefore estimate the above nested model to let the data determine the appropriate value of s.

Because our model is non-linear, we cannot use the differenced panel estimators for cross-country growth regressions that have become popular in the literature

[e.g. Caselli, Esquivel and Lefort (1996), Easterly, Loayza and Montiel (1997), and Benhabib and Spiegel (2000)]. Instead, we estimate the nested specification above in a cross-sectional sample of long-term growth using maximum likelihood. In order to minimize problems with endogeneity, we use initial values for human capital stocks and initial total factor productivity. As we are comparing these initial values to the nations' subsequent growth experiences over the next 35 years, endogeneity issues are unlikely to be a problem.

We also conduct a number of robustness checks. First, there is a concern about the quality of initial human capital values as a proxy of the human capital stock available over the estimation period. Recall that our specification implies that human capital is a measure of a nation's capacity to conduct innovation activity (accounted by the first term in the specification), and technology adoption from abroad (captured by the second term in the specification). However, many of the nations in our sample exhibited dramatic growth in their human capital stocks over this period, as measured by average years of schooling. A number of nations, including Nepal, Togo, Iran, Ghana, Syria, and the Central African Republic, actually had more than a five-fold increase in their average years of schooling in the population over the age of 25. This implies that the initial stocks of human capital in 1960 may poorly represent the stocks of human capital available to a nation later on in the sample period. We therefore also report results using average human capital levels over the estimation period.[21] However, this measure is likely to suffer more from endogeneity issues than initial human capital levels, as a nation's financial ability to increase the average human capital levels of its citizens is likely to be increasing in its rate of output and total factor productivity growth. Fortunately, as we demonstrate below, our results are fairly robust to either measure of the stock of human capital.

[21] Average human capital levels are calculated as the simple averages of beginning (1960) and ending (1995) human capital levels.

Second, since we are estimating a cross-section, we are unable to condition on country-specific fixed effects. In response, we further examine the robustness of our results to the introduction of a number of conditioning variables. Using data obtained from Sachs and Warner (1997), we introduce a number of geo-political characteristics, including a Sub-Saharan Africa dummy, a dummy for countries that are not landlocked, a dummy for tropical countries, a dummy for initial life expectancy, a dummy for ethnolinguistic fractionalization, and a dummy for openness over the estimation period.

5.3. Results

5.3.1. Base specification

Our results with initial stocks of human capital are shown in Table 2. Our base specification is reported in Model 1. It can be seen that the coefficient on human capital, which represents $(g + c/s)$ in the specification above, enters significantly with a positive coefficient in log levels at a 5 percent confidence level, consistent with the notion of human capital as a facilitator of own innovation predicted by the theory. The next term represents the coefficient on the catch-up term, -(c/s) in the above specification. This term enters as predicted with a negative and statistically significant sign at a five-percent confidence level. Finally, our point estimate of s is equal to 2.304. This number is not significantly different from 1, but is significantly greater than 0 at a ten percent confidence level. These results therefore favor the logistic specification, suggesting that there is some initial human capital level below which a country would fall farther and farther behind the leader national in total factor productivity over time. We investigate this possibility in more detail below.

One disappointing result in our base specification is that our point estimate for human capital lies below that of the catch-up term in absolute value. This implies

that our point estimate for g is negative, which is implausible. However, this point estimate is insignificantly different from 0 and does include positive values for any standard confidence level. Nevertheless, the negative point estimate does become a problem for our data exploration. In particular, using the negative point estimate for g precludes the existence of a positive critical human capital stock below which catch up in total factor productivity cannot occur.

As discussed above, the problem with the specification of Model 1 is that our theory does not call for for the a constant term independent of human capital to account for total factor productivity growth. Consequently, Model 2 repeats our base specification with the constant term excluded. It can be seen that our qualitative results are robust to the exclusion of a constant term. Human capital in log levels again enters significantly with a positive coefficient at a 5 percent confidence level, while the catch-up term is again significantly negative at a 5 percent confidence level, as predicted by the theory. Our point estimate of s is a little higher, at 3.164, but as before we cannot reject the hypothesis that s is equal to 1 at standard confidence levels, although we again reject the hypothesis that s is less than or equal to 0 at a 10 percent confidence level. Moreover, it can be seen that our point estimate for g is positive with this specification, allowing us to calculate a critical human capital stock below which catch-up in growth rates will not occur.

Models 3 and 4 repeat our estimation with and without a constant term, with s constrained to equal 1. This results in a linear specification and provides a robustness check of the coefficients obtained in our non-linear specification. It can be seen that our point and standard error estimates are very close to those obtained with s unconstrained. Both with and without a constant term, human capital enters significantly with a positive coefficient in log levels at a 5 percent confidence level. Moreover, the catch-up term coefficient is again negative and significant at a 5 percent confidence level, as predicted. These results suggest

that our findings are not dependent on the non-linear estimation of s to obtain coefficient estimates consistent with the notion of human capital playing a positive role in facilitating both innovation and catch-up.

5.3.2. Average human capital levels

Our first set of robustness checks repeats our estimation using average levels of human capital over the estimation period rather than initial human capital values.[22] As discussed above, we do this to address the concern some nations' stocks of human capital changed dramatically over the estimation period, and therefore that initial human capital values may be relatively noisy indicators of the average levels of human capital over the estimation period that determined their TFP growth .

The results incorporating this change are shown in Table 3. It can be seen that our qualitative results are fairly robust. Average human capital levels enter positively and significantly, as predicted, at a 5 percent confidence level, as do the coefficient estimates for the catch-up term. The magnitudes of these coefficients are similar to those obtained with initial human capital stocks, but they are both somewhat larger in absolute value. This increase is interesting because average measured human capital levels are larger than initial human capital levels, as all nations experienced some increase in average years of schooling over the estimation period.

Our estimates of s in Models 1 and 2 are very close to 1, which would again favor our logistic specification, but the large standard errors associated with our estimates of s leave it insignificantly different from 0 at standard confidence levels.

[22] Average stocks are estimated using simple averages of period beginning and ending values.

5.3.3. Conditioning on Country Characteristics

Because we are estimating a cross-section, we obviously are precluded from using panel estimators, such as country fixed and random effects, to control for differences in country characteristics outside of our theory that may independently influence total factor productivity growth. To account for these other possible influences, we introduce a number of conditioning variables into our specification from the Sachs and Warner (1997) data set.[23] The conditioning variables introduced are *Sub-Sahara,* a dummy indicating Sub-Saharan African nations, *Landlocked,* a dummy indicating a nation lacking navigable access to the sea, *Tropics,* a variable measuring the share of land area subject to a tropical climate, *Life,* the log of life expectancy at birth measured between 1965 and 1970, *Ethling,* a measure of ethno-linguistic fractionalization, and *Openness,* an indicator of the degree to which domestic policy favors free trade.

We first present our results with all of the conditioning variables included, and then sequentially drop the *Sub-Sahara* and *Openness* variables. Our results are shown in Table 4. Note that the inclusion of these conditioning variables reduces our sample size from 84 to 75 countries. Models 1 and 2 report our results for our base specifications with all of the conditioning variables included. It can be seen that human capital in log levels is not positive at a statistically significant level in either specification. This result is attributable more to a substantial increase in our standard error estimate rather than a change in the point estimate of the coefficient, which does not change much in value. On the other hand, it appears that the catch-up term is robust to the inclusion of these conditioning variables, as it enters significantly with a negative coefficient at a five percent confidence level, as predicted. Finally, our point estimates of s are still close to 1. We cannot reject that s is negative at standard confidence levels when our intercept

[23] See Sachs and Warner (1997) for original data sources.

term is included, but we can with it excluded (Model 2).[24]

Models 3 and 4 omit the *Sub-Sahara* dummy. It can be seen that human capital in log levels is still insignificant when the constant term is included, but is now significant at a 10 percent confidence level when the constant term is excluded. The catch-up term is still significantly negative at a 5 percent confidence level, as predicted. Our point estimates for s are still close to 1, with s entering significantly with a greater than zero coefficient at a 10 percent confidence level with and without the inclusion of a constant term.

Finally, Models 5 and 6 omit the *Openness* variable. Human capital in log levels is insignificant with the constant term included, but is positive and significant, as predicted, with the exclusion of the constant term at a 10 percent confidence level. The catch-up term is still significantly negative at a 5 percent confidence level, as predicted. Our point estimates for s are again close to 1, although s is insignificantly different from zero both with and without the inclusion of a constant term in our specification.

In summary, it appears that the catch-up term is strongly robust to the inclusion of the conditioning variables, while the estimates of s are still close to one, but of mixed significance. It would therefore be fair to characterize these coefficient estimates to be fairly robust to the inclusion of the conditioning variables.[25]

[24] To determine whether the differences here were attributable to the inclusion of the conditioning variables or the reduction in sample size, we estimated our models with the smaller 75 country sample reported here with the conditioning variables excluded. We obtained similar results to those in the larger sample. In particular, we obtained a positive and significant coefficient on human capital in log levels. This indicates that the differences in results reported here are attributable to the inclusion of the conditioning variables.

[25] To investigate the possibility that technological catch-up was facilitated by other variables than human capital, we substituted our *Life* and *Openness* conditioning variables for human capital in our base specification. The estimate for s was positive, but insignificant in all specifications. The coefficients on *Life,* both on their own and interacted with backwardness,

However, human capital in log levels was somewhat less robust. This result may not be surprising for a number of reasons: First, the conditioning variables, such as initial life expectancy and subsequent openness, are likely to be correlated with initial human capital levels. Indeed, initial life expectancy may be considered to be an alternative indicator of investment in human capital for many developing countries. Second, Benhabib and Spiegel (1994) found that initial human capital, which determines the rate of own-country innovation, was unimportant for a sub-sample of poorer developing countries. The introduction of our conditioning may have exposed the relatively weak role that innovation plays in total factor productivity growth for the poorer nations in our sample.[26]

---

were consistent with the theory and significant with the constant term included, but insigifnicant with it excluded. The coefficients on *Openness,* however, both on their own and interacted with backwardness, were very insignificant. As a whole, this exercise provided weak evidence of robustness for the logistic specification. Yet the imprecision of our measurements and the high correlation between country characteristic measures makes it difficult to evaluate the precise contribution of human capital relative to other potential institutuional characteristics that can facilitate catch-up. For example, the correlation coefficient between h_{i60} and $Life$ is 0.85. These results are available from the authors on request.

[26] We also examined the robustness of our results to splitting the sample with the conditioning variables included. We split the sample into OECD and non-OECD nations. Our coefficient values for both sub-samples were of the correct sign and significant. However, we also found that the point estimate of the innovation term was larger for the OECD sub-sample, while that for the catch-up term was larger in absolute value for the non-OECD sample. This supports our findings in Benhabib and Spiegel (1994) that innovation is more important for the developed countries, while catch-up is more important for the developing nations. These regression results are also available upon request from the authors.

6. Model Prediction

6.1. Model Forecasting

Given a nation's initial values of H_{i60} and $B_i(60)$, our transition equation 4.4 gives us a predicted value of B at the end of our sample in 1995. Figure 2 displays the predicted values of $B_i(95)$ conditional on H_{i60} and $B_i(60)$. One can see the logistic "s" form, consistent with a logistic model of technology diffusion, of our predicted values from our estimation above. Countries which have both low initial total factor productivity relative to the leader and low levels of human capital are in the low-growth portion of the plane: their predicted 1995 total factor productivity levels relative to the leader lie close to, or even below, their 1960 values. There is then a rapid acceleration in the middle range, tapering off as nations approach the total factor productivity levels of the leader.

We show both the actual realizations and the predictons of our model in Figure 3. Expected values of $B_i(95)$ for the nations in our sample based on equation 4.4 are plotted against their realized values in 1995. The model does a fairly good job of predicting relative future productivity levels. As a measure of our goodness of fit, we calculated the coefficient of determination of the model. The ratio of residual sum-of-squared errors to the variation in the sample was only, 0.115, which would correspond to an R-squared of 88.5 percent.

However, there does appear to be some systematic errors in our forrecasts. In particular, we seem to be systematically overestimating relative total factor productivity growth for the least backward, highest initial productivity countries like the Asian Tigers, so that the residuals for these countries are nearly all negative. This result, which suggests an even more prounced 's' curve, is puzzling, but appears to leave room for future refinements in our theory.

6.2. Negative Catch-up Countries

A more qualitative metric of the quality of fit of our model is how well it makes the discrete prediction of whether coutries will be on a positive catch-up path or not. The theory above suggests that below a certain threshold level of human capital, relative to the leader nation, a country could find its total factor productivity growth sufficiently slow that it would not exhibit convergence in total factor productivity, but would instead fall farther and farther behind the leader nation over time. In particular, we can re-write the "catch-up condition" in equation 4.7 as

$$H_{it}^{*} = \exp\left(\frac{sgh_{mt}}{sg+c}\right) \tag{6.1}$$

where h_{mt} represents the log of human capital in the leader nation at time t. Countries that find themselves with human capital stocks below H_{it}^{*} will experience total factor productivity growth at a slower pace than the leader country.

Table 5 shows the point estimates for g, c, and s based on our estimation results for models 1 through 4 in Tables 2 and 3. As we discussed above, we cannot calculate a critical human capital stock for Model 1 in Table 2 because of our negative point estimate for g. Consequently, we concentrate on the point estimates obtained in Model 2 of Table 2, where the specification excludes a constant term independent of human capital. As we show below, our estimates of the critical human capital stocks are similar for all of our models.

With the United States as our leader in total factor productivity, the point estimates obtained with Model 2 indicate that countries with average schooling in the population over the age of 25 below 1.78 years will display slower total factor productivity growth than the leader nation. We note that the critical human capital stocks were relatively insensitive to model specification or the use of initial or average human capital levels.

Similarly, we can also calculate the average years of schooling in the population needed to experience faster total factor productivity growth than the United States

in 1995. Because of the increase in average years of schooling in the United States, the point estimates for H^*_{i95} are uniformly larger than those for H^*_{i60}. Again using our point estimates from Model 2, we estimate the critical level of average years of schooling in the population to be 1.95. This increase in the threshold level of human capital is due to the fact that with a larger stock of human capital, the leader nation will be innovating at a faster pace. Consequently, other nations will need to exhibit a faster pace of catch-up to experience faster total factor productivity growth than the leader.

We use these estimated critical human capital stocks to conduct 2 explorations in the data. First, we can identify nations in our sample that would be predicted to exhibit slower growth in total factor productivity than the United States in 1960. This would include all nations with human capital levels in 1960 below 1.78 years of schooling. Our results are shown in Table 2. Based on our point estimates, we identify 27 nations as being below the critical human capital stock level in 1965. These nations are listed in Table 6, along with their average initial human capital stock levels.

The second column examines the growth performance of these nations over the subsequent 35 years in our sample. While it is not a formal test of our model, it is rather striking that 22 of the 27 nations predicted to exhibit slower total factor productivity growth than the United States actually did so over the course of our sample. This is markedly different than the overall sample share, where 49 of the 84 countries exhibited faster total factor productivity growth than the United States. Consequently, the subsequent performance of these nations appears to support the possibility of a logistic form of technology diffusion.

Our second data exploration concerns the question of whether there are any nations that are still below the critical human capital stock, so that they are expected to have slower total factor productivity growth than the United States in the future. We investigate this question using our 1995 data. As mentioned

above, the critical human capital stock using any model specification is estimated to have increased slightly between 1960 and 1995, from 1.78 average years of schooling to 1.95. Nevertheless, the good news is that because many developing nations have made substantial efforts to increase primary education rates in their populations, there are few countries who failed to meet this criterion in 1995.

The four nations that fell below the critical human capital level in 1995 are listed in Table 7. They are Mali, Niger, Mozambique and Nepal. While the success of the rest of the world in acquiring sufficient human capital to be on positive catch-up path in total factor productivity is reassuring, the situation faced by these four nations is still alarming. As shown in Table 7, none of these nations has a total factor productivity level exceeding 15 percent of that in the United States. In contrast, the average ratio of the total factor productivity of a nation in our sample to that of the United States is approximately 44 percent. Our model therefore predicts that these nations will remain notably poor in the absence some sort of policy intervention.

7. Conclusion

This paper generalizes the Nelson-Phelps catch-up model of technology diffusion facilitated by levels of human capital. We allow for the possibility that the pattern of technology diffusion is exponential. This specification predicts that nations will exhibit positive catch-up in growth rates. In contrast a logistic diffusion specification implies that a country with a sufficiently small capital stock may exhibit slower total factor productivity growth than the leader nation.We then derive a nonlinear specification for total factor productivity growth that nests these two specifications. We test this specification for a cross-section of 84 countries. Our results favor the logistic specification over the exponential, and other estimated parameters are consistent with our theoretical predictions. The

catch-up term in our specification is robust to a number of sensitivity checks, including the use of average rather than initial levels of human capital and the inclusion of a variety of geo-political conditioning variables commonly used in the literature. This supports the notion that human capital plays a positive role in the determination of total factor productivity growth rates through its influence on the rate of catch-up. However, the direct performance of the human capital term on its own is somewhat less robust.

Using the coefficient estimates from our parametric estimation, we then calculate the critical human capital stocks needed to achieve positive total factor productivity growth in 1960 and 1995. Our results identify 27 nations as falling below the critical human capital level in 1960, while only 4 nations remain below the critical human capital level in 1995.

The historic experiences of these nations support our theory well. 22 of the 27 nations predicted to have slower growth than the leader nation (the United States) actually did so over the subsequent 35 years. This contrasts markedly with the overall experience of the nations in our sample, where 49 of the 84 nations experienced faster total factor productivity growth than the leader nation.

References

[1] Acemoglu, Daron (2002), "Factor Prices and Technical Change," in Philippe Aghion ed., *Festschrift in Honor of Edmund Phelps,* forthcoming.

[2] Acemoglu, Daron, Aghion, Philippe and Fabrizio Zilibotti, (2002), "Distance to Frontier, Selection and Economic Growth," NBER Working Paper no. w9066, 2002.

[3] Aghion, Philippe, Christopher Harris and Johnathan Vickers (1997), "Competition and Growth with Step-by-Step Innovation: An Example," *European Economic Review,* April 1997, 41(3-5), 771-82

[4] Banks, Robert B. (1994), *Growth and Diffusion Phenomena,* (Springer Verlag, Berlin).

[5] Barro, Robert J. and Jong Wha Lee. (1993), "International Comparisons of Educational Attainment," *Journal of Monetary Economics,* 32, 363-394.

[6] Barro, Robert J., and Xavier Sala-i-Martin, (1995), *Economic Growth,*

[7] Barro, Robert J., and Xavier Sala-i-Martin, (1997), "-Technological Diffusion, Convergence and Growth," *Journal of Economic Growth,* 1, 1-26

[8] Bartel, Ann P, and Frank R. Lichtenberg, (1987), "The Comparative Advantage of Educated Workers in Implementing New Technology," *Review of Economics and Statistics,* February, 69(1), 1-11.

[9] Basu, Susanto, and David N. Weil, (1998), "Appropriate Technology and Growth," *Quarterly Journal of Economics,* November, 113(4), 1025-54.

[10] Benhabib, Jess and Mark M. Spiegel. (1994), "The Role of Human Capital in Economic Development: Evidence from Aggregate Cross-Country Data," *Journal of Monetary Economics,* 34, 143-173.

[11] Benhabib, Jess and Mark M. Spiegel, (2000), "The Role of Financial Development in Growth and Investment," *Journal of Economic Growth,* 5, 341-360.

[12] Bils, Mark, and Peter J. Klenow, (2000), "Does Schooling Cause Growth?," *American Economic Review,* December, 90(5), 1160-83.

[13] Branstetter, Lee G., (2001), "Are Knowledge Spillovers International or Intranational in Scope? Microeconometric Evidence from the U.S. and Japan," *Journal of International Economics,* February, 53(1), 53-79.

[14] Caselli, Francesco, Gerardo Esquivel, and Fernando Lefort. (1996), "Reopening the Convergence Debate: A New Look at Cross-Country Growth Empirics," *Journal of Economic Growth,* 1, 363-390.

[15] Coe, David T. and Elhanan, (1995), "International R&D Spillovers," *European Economic Review,* May, 39(5), 859-87

[16] Durlauf, Steven, N. and Paul A. Johnson, (1995), "Multiple Regimes and Cross-Country Growth Behavior," *Journal of Applied Econometrics, 365-384.*

[17] Duffy, Johnathan and Christopher Papageorgiou, (2000), "A Cross-Country Empirical Investigation of the Aggregate Production Function Specification," *Journal of Economic Growth,* 5, 87-120.

[18] Eaton, Jonathan ans Samuel Kortum, (1996), "Trade in Ideas: Patenting and Productivity Growth in the OECD," *Journal of International Economics,* 40, 251-278.

[19] Eaton, Jonathan ans Samuel Kortum, (1996), "International Technology Diffusion: Theory and Measurement" *International Economic Review,* 40, 537-570.

[20] Eeckhout, Jan and Boyan Jovanovic, (2000), "Knowledge Spillovers and Inequality." *American Economic Review* 92, no. 5.

[21] Easterly, William, Norman Loayza, and Peter Montiel. (1997), "Has Latin America's Post-Reform Growth Been Disappointing?," *Journal of International Economics,* 43, 287-311.

[22] Foster, Andrew D. and Mark R. Rosenzweig, (1995), "Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture," *Journal of Political Economy,* December, 103(6), 1176-1209.

[23] Gerschenkron, A., (1962), Economic backwardness in historical perspective, Cambridge, Belknap Press of Harvard University Press.

[24] Griliches, Zvi, (1992), "The Search for R&D Spillovers," *Scandinavian Journal of Economics,* 94, 29-47.

[25] Grossman, Gene M. and Elhanan Helpman, (1991), "Trade, Knowledge Spillovers and Growth," *European Economic Review,* 35, 517-26.

[26] Hall, Robert E. and Charles I. Jones, (1999), "Why Do Some Countries Produce So Much More Output Per Worker Than Others?," *Quarterly Journal of Economics,* February, 114(1), 83-116

[27] Hanushek, Eric A. and Dennis D. Kimko, (2000), "Schooling, Labor-Force Quality, and the Growth of Nations," *American Economic Review,* December, 90(5), 1184-1208

[28] Helpman, Elhanan, (1993), "Innovation, Imitation, and Intellectual Property Rights," *Econometrica,* 61, 1247-80.

[29] Howitt, Peter and David Mayer-Foulkes, (2002) "R&D, Implementation and Stagnation: A Schumpeterian Theory of Convergence Clubs," NBER Working Paper Series, no. 9104, Cambridge.

[30] Islam, Nazrul. (1995). "Growth Empirics: A Panel Data Approach," *Quarterly Journal of Economics,* 110, 1127-1170.

[31] Kyriacou, George, (1991), "Level and Growth Effects of Human Capital," C.V. Starr Center Working Paper no. 91-26.

[32] Klenow, Peter J., and Andrés Rodriguez-Clare, (1997), "The Neoclassical Revival in Growth Economics: Has It Gone Too Far?," *N.B.E.R. Macroeconomics Annual,* 73-103.

[33] Krueger, Alan B. and Mikael Lindahl, (2001), "Education for Growth: Why and for Whom?," *Journal of Economic Literature,* December, 39(4), 1101-36

[34] Mansfield, Edwin, (1968), *Industrial Research and Technogical Innovation*, Norton, New York.

[35] Nadiri, M. Ishaq, and Seongjun Kim, (1996), "International R&D Spillovers, Trade and Productivity in Major OECD Countries," N.B.E.R. Working Paper no. 5801, October.

[36] Nelson, Richard R. and Edmund S. Phelps (1966), "Investment in Humans, Technological Diffusion, and Economic Growth," *American Economic Review,* 56, 69-75.

[37] Parente, Stephen L.., and E. C. Prescott, "Barriers to Technology Adoption and Development," Journal of Political Economy, 102 (1994), 298-321,

[38] Pritchett, L. (1996), "Where has all the education gone?", World Bank Policy Research Working Paper #1581

[39] Psacharopoulos, George, and Ana Maria Arriagada (1986), "The Educational Attainment of the Labor Force: An International Comparison," *International Labor Review,* 125(5), 561-574.

[40] Richards, F.J., (1959), "A Flexible Growth Function for Empirical Use," *Journal of Experimental Botany*, 290-300

[41] Romer, P., (1990), "Endogenous Technical Change," *Journal of Political Economy,* 98, S71-S102.

[42] Sachs, Jeffrey. D. and Andrew M. Warner, (1997), "Fundamental Sources of Long-Run Growth,". *American Economic Review*, May, 87(2), 184-188.

[43] Segerstrom, P., (1991), "Innovation, Imitation, and Economic Growth," *Journal of Political Economy,* 94, 1163-1190.

[44] Sharif and Ramanthran (1981), "Binomial Innovation Diffusion Models with Dynamic Potential Adopter Population, Technological Forecasting and Social Change 20, 63-87.

[45] Temple, Jonathan, "The New Growth Evidence," Journal of Economic Literature, March, 37(1), 112-56.

[46] Welch, Finis, (1975), "Human Capital Theory: Education, Discrimination, and Life Cycles," *American Economic Review,* May, 65(2), 63-73.

Table 1

Total Factor Productivity Estimates (1960-1995)

Country	log TFP_{1960}	log TFP_{1995}	avg annual log growth of TFP (1960-1995)
Mozambique	0.5010	-0.0353	-0.0153
Niger	0.2045	-0.2983	-0.0144
Central African Rep.	0.4180	-0.0791	-0.0142
Nicaragua	0.4487	0.0546	-0.0113
Zambia	-0.2912	-0.5857	-0.0084
Mali	-0.1092	-0.2677	-0.0045
Senegal	0.3209	0.1634	-0.0045
Venezuela	1.0141	0.9306	-0.0024
Togo	-0.1249	-0.1917	-0.0019
Cameroon	0.3181	0.2649	-0.0015
Tanzania	-1.0572	-1.0181	0.0011
Bolivia	0.3817	0.4642	0.0024
Honduras	0.1513	0.2597	0.0031
El Salvador	0.7495	0.8820	0.0038
Guyana	0.0168	0.1989	0.0052
Peru	0.4039	0.6054	0.0058
Argentina	0.9538	1.1675	0.0061
Uganda	0.0519	0.2721	0.0063
South Africa	0.8463	1.0689	0.0064
Jamaica	0.2297	0.4554	0.0064
Philippines	0.2176	0.4506	0.0067
Costa Rica	0.6131	0.8480	0.0067
Bangladesh	-0.0997	0.1442	0.0070
Jordan	0.4289	0.6773	0.0071
New Zealand	1.1840	1.4505	0.0076
Uruguay	0.8978	1.1733	0.0079
Nepal	-0.3250	-0.0416	0.0081
Malawi	-0.7672	-0.4742	0.0084
Algeria	0.3615	0.6622	0.0086
Ghana	-0.2121	0.0893	0.0086
Guatemala	0.5197	0.8215	0.0086
Switzerland	1.2467	1.5526	0.0087
Kenya	-0.2842	0.0390	0.0092

Country	log TFP_{1960}	log TFP_{1995}	avg annual log growth of TFP (1960-1995)
Mexico	0.6282	0.9549	0.0093
Papua New Guinea	0.3175	0.6532	0.0096
Iran	0.3787	0.7390	0.0103
Lesotho	-0.4715	-0.1054	0.0105
Trinidad &Tobago	0.8535	1.2695	0.0119
Fiji	0.3940	0.8118	0.0119
Ecuador	0.1191	0.5526	0.0124
Sweden	1.0855	1.5350	0.0128
Dominican Rep.	0.2220	0.6859	0.0133
United Kingdom	1.1090	1.5778	0.0134
Canada	1.1711	1.6541	0.0138
Australia	1.1472	1.6339	0.0139
Denmark	1.1227	1.6215	0.0143
Paraguay	0.4728	0.9894	0.0148
Turkey	0.4371	0.9546	0.0148
Colombia	0.4648	0.9855	0.0149
Netherlands	1.0327	1.5617	0.0151
Zimbabwe	-0.2344	0.2948	0.0151
United States	1.3257	1.8626	0.0153
Sri Lanka	0.0648	0.6074	0.0155
Finland	0.8676	1.4237	0.0159
Iceland	0.9602	1.5301	0.0163
Chile	0.6381	1.2141	0.0165
India	-0.2360	0.3458	0.0166
Panama	0.2486	0.8324	0.0167
France	0.9176	1.5088	0.0169
Ireland	0.8202	1.6031	0.0182
Belgium	0.9147	1.5555	0.0183
Syria	0.1391	0.7957	0.0188
Brazil	0.2618	0.9204	0.0188
Greece	0.5097	1.1877	0.0194
Austria	0.8583	1.5445	0.0196
Norway	0.8808	1.5879	0.0202
Italy	0.8291	1.5379	0.0202
Israel	0.7494	1.4757	0.0163
Pakistan	-0.4390	0.3175	0.0216

Country	log TFP_{1960}	log TFP_{1995}	avg annual log growth of TFP (1960-1995)
Spain	0.6153	1.4203	0.0230
Mauritius	0.6394	1.4829	0.0241
Portugal	0.4739	1.3254	0.0243
Indonesia	-0.1621	0.7056	0.0248
Barbados	0.5475	1.4540	0.0259
Malaysia	0.2549	1.1852	0.0266
Romania	-0.3987	0.5327	0.0266
Japan	0.5632	1.5851	0.0292
Botswana	-0.1326	0.9935	0.0322
Cyprus	0.3582	1.5217	0.0332
Thailand	-0.3058	0.9102	0.0347
Hong Kong	0.4578	1.8604	0.0401
Rep. of Korea	-0.0429	1.3646	0.0402
Singapore	0.1202	1.6285	0.0431
Rep. of China, Taiwan	0.1046	1.6140	0.0431

Table 2

Regression Results: Log H_{1960}

	Model 1	Model 2	Model 3	Model 4
C	0.0083**	–	0.0085**	–
	(0.0016)		(0.0016)	
$\ln(H_{1960})$	0.0080**	0.0116**	0.0100**	0.0134**
	(0.0019)	(0.0016)	(0.0023)	(0.0025)
$\ln(H_{1960}) * \left(\frac{TFP_i}{TFP_m}\right)^s$	-0.0086**	-0.0085**	-0.0089**	-0.0072**
	(0.0032)	(0.0039)	(0.0036)	(0.0025)
s	2.304*	3.164*	1	1
	(1.405)	(1.892)		
# of observations	84	84	84	84
log likelihood	264.5	252.4	263.9	263.9
Wald P-value	0.00	0.00	0.00	0.00

note: Estimation by maximum likelihood with standard errors presented in parentheses. ** denotes statistical significance at the 5% confidence level while * denotes statistical significane at the 10% confidence level.

Table 3

Regression Results: Log $\overline{H}_{1960-1995}$

	Model 1	Model 2	Model 3	Model 4
C	-0.0030	–	-0.0030	–
	(0.0024)		(0.0024)	
$\ln(\overline{H}_{1960-1995})$	0.0175**	0.0150**	0.0184**	0.0159**
	(0.0046)	(0.0039)	(0.0026)	(0.0017)
$\ln(\overline{H}_{1960-1995}) * \left(\frac{TFP_i}{TFP_m}\right)^s$	-0.0129**	-0.0116**	-0.0135**	-0.0122**
	(0.0039)	(0.0036)	(0.0031)	(0.0029)
s	1.151	1.192	1	1
	(0.783)	(0.862)		
# of observations	84	84	84	84
log likelihood	274.5	273.7	274.4	273.6
Wald P-value	0.00	0.00	0.00	0.00

note: Estimation by maximum likelihood with standard errors presented in parentheses. ** denotes statistical significance at the 5% confidence level while * denotes statistical significane at the 10% confidence level.

Table 4

Regression Results: Log $\overline{H}_{1960-1995}$ and Geo-Political Variables

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
C	-0.0671	–	-0.0778*	–	-0.1394**	–
	(0.0463)		(0.0461)		(0.0488)	
$\ln(\overline{H}_{1960-1995})$	0.0077	0.0070	0.0072	0.0067*	0.0092	0.0080**
	(0.0061)	(0.0043)	(0.0054)	(0.0038)	(0.0077)	(0.0038)
$\ln(\overline{H}_{1960-1995})$	-0.0196**	-0.0164**	-0.0194**	-0.0159**	-0.0213**	-0.0142**
$*\left(\frac{TFP_i}{TFP_m}\right)^s$	(0.0066)	(0.0045)	(0.0059)	(0.0041)	(0.0082)	(0.0043)
s	0.9302	1.1380*	0.9866*	1.2250*	0.8375	1.2780
	(0.5796)	(0.6534)	(0.5621)	(0.6414)	(0.5857)	(0.7953)
ssafrica	-0.0041	-0.0049*	–	–	-0.0047	-0.0065**
	(0.0030)	(0.0030)			(0.0032)	(0.0033)
access	-0.0018	-0.0027	-0.0026	-0.0038	0.0002	-0.0015
	(0.0027)	(0.0026)	(0.0026)	(0.0026)	(0.0029)	(0.0030)
tropics	-0.0070**	-0.0074**	-0.0073**	-0.0078**	-0.0086**	-0.0096**
	(0.0026)	(0.0026)	(0.0026)	(0.0026)	(0.0027)	(0.0028)
life1	0.0201*	0.0031**	0.0228*	0.0031**	0.0393**	0.0044**
	(0.0117)	(0.0007)	(0.0117)	(0.0007)	(0.0123)	(0.0008)
ethling	0.0001	0.0000	0.0000	-0.0000	0.0001*	0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
openess	0.0128**	0.0140**	0.0128**	0.0143**	–	–
	(0.0027)	(0.0026)	(0.0027)	(0.0026)		
# of observations	75	75	75	75	78	78
log likelihood	259.8	258.7	258.8	257.4	261.2	257.3
Wald P-value	0.00	0.00	0.00	0.00	0.00	0.00

note: Estimation by maximum likelihood with standard errors presented in parentheses. ** denotes statistical significance at the 5% confidence level while * denotes statistical significane at the 10% confidence level. See text for definitions of the conditioning variables.

Table 5

Point Estimates

H_{1960}

	Model 1	Model 2	Model 3	Model 4
g	-0.0006	0.0031	0.0012	0.0063
c	0.0198	0.0268	0.0089	0.0072
s	2.3040	3.1645	1	1
H^*_{60}	n.a.	1.78	1.29	2.75
H^*_{95}	n.a.	1.95	1.35	3.22

$\overline{H}_{1960-1995}$

	Model 1	Model 2	Model 3	Model 4
g	0.0046	0.0034	0.0049	0.0037
c	0.0149	0.0138	0.0135	0.0122
s	1.1515	1.1921	1	1
H^*_{60}	1.76	1.63	1.78	1.65
H^*_{95}	1.93	1.76	1.95	1.79

note: g, c, and s are obtained from the point estimates presented in Tables 2 and 3. H^*_{60} and H^*_{95} represent the minimal initial estimated stock of human capital needed for positive predicted growth relative to the leader nation.

Table 6

Nations with Slow TFP Growth (1960)

Country	H_{1960}	$(\text{TFP Growth}_i) - (\text{TFP Growth}_{USA})$
Nepal	0.07	-0.0072
Mali	0.17	-0.0199
Niger	0.20	-0.0297
Mozambique	0.26	-0.0307
Togo	0.32	-0.0172
Central African Republic	0.39	-0.0295
Iran	0.63	-0.0050
Pakistan	0.63	0.0063
Ghana	0.69	-0.0067
Bangladesh	0.79	-0.0084
Algeria	0.97	-0.0067
Syria	0.99	0.0034
Uganda	1.10	-0.0090
Indonesia	1.11	0.0095
Papua New Guinea	1.13	-0.0057
Kenya	1.20	-0.0061
Cameroon	1.37	-0.0169
Jordan	1.40	-0.0082
Guatemala	1.43	-0.0067
India	1.45	0.0013
Botswana	1.46	0.0168
Zimbabwe	1.54	-0.0002
Senegal	1.60	-0.0198
Zambia	1.60	-0.0238
Honduras	1.69	-0.0122
Malawi	1.70	-0.0070
El Salvador	1.70	-0.0116

note: The nations listed are those with 1960 human capital levels below 1.78, the minimum needed for TFP catchup according to Model 2 in Table 2.

Table 7

Nations with Slow TFP Growth (1995)

Country	H_{1995}	$\frac{TFP_{1995i}}{TFP_{1995USA}}$
Mali	0.69	0.1188
Niger	0.69	0.1152
Mozambique	1.01	0.1499
Nepal	1.53	0.1489

note: The nations listed are those with 1995 human capital levels below 1.95, the minimum needed for TFP catchup according to Model 2 in Table 2. For the full 84 country sample, $\overline{\frac{TFP_{1995i}}{TFP_{1995USA}}} = 0.4377$

Figure 1

TFP Growth vs Initial Human Capital

Files/Qualcomm/Eudora/attach/growth11-Figure1.wmf

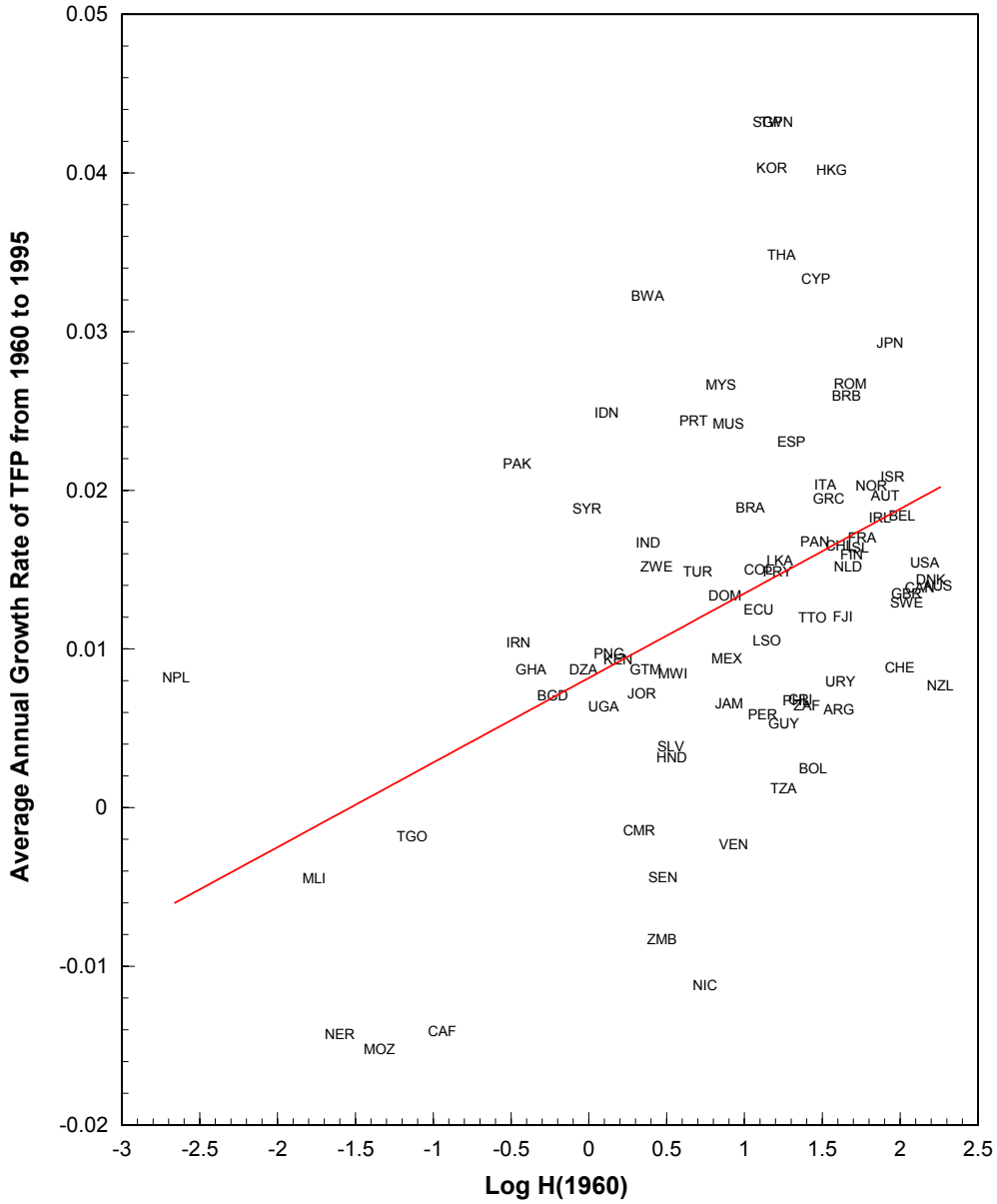
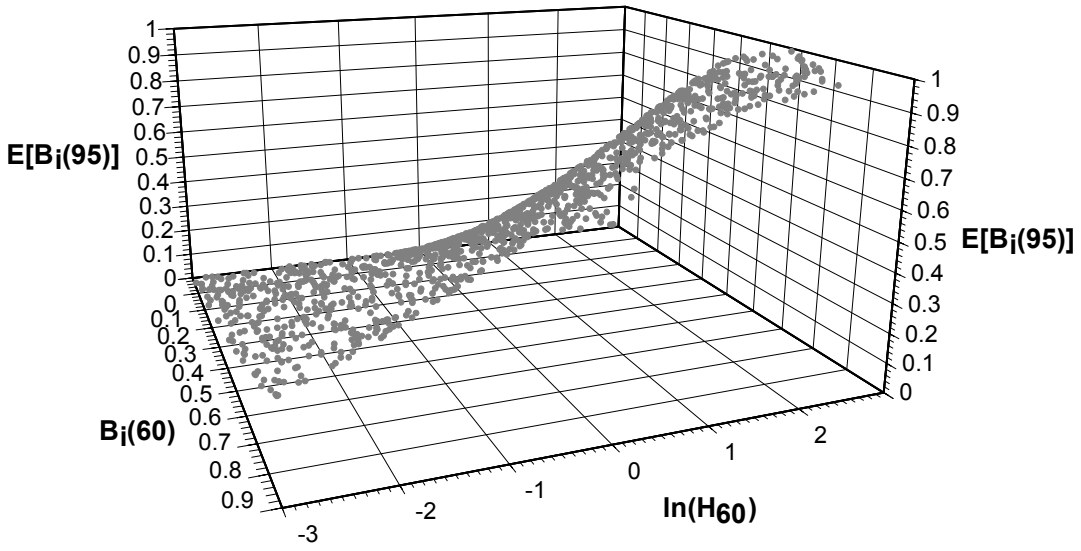


Figure 2

Predicted Values of $B_i(1995)$[1]

Files/Qualcomm/Eudora/attach/growth11-Figure2.wmf

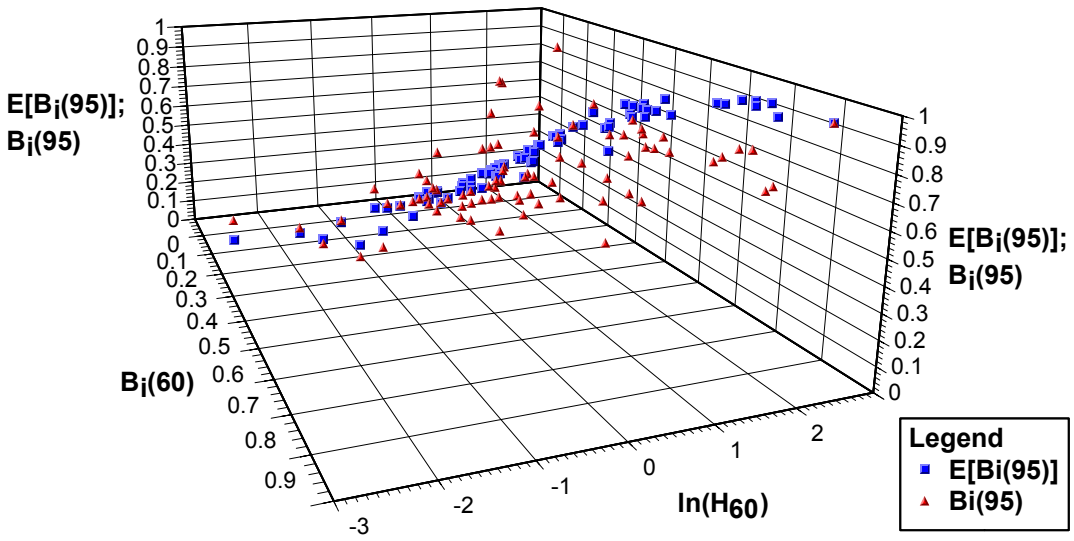


[1]Predicted values of $B_i(95)$are based on initial backwardness in TFP, $B_i(60)$, and the log of initial stock of human capital. $B_i(t)$represents the ratio of TFP in country $i$to TFP in the leader country (United States) at time t. The sample encompasses the entire range of values for backwardness and human capital.

Figure 3

Predicted and Actual Values of $B_i(1995)$[1]

Files/Qualcomm/Eudora/attach/growth11-Figure3.wmf



[1]Predicted values of $B_i(95)$are based on initial backwardness in TFP, $B_i(60)$, and the log of initial stock of human capital. $B_i(t)$represents the ratio of TFP in country $i$to TFP in the leader country (United States) at time t. The sample includes observed data points only.

GROWTH STRATEGIES*

Dani Rodrik
Harvard University

John F. Kennedy School of Government
79 Kennedy Street
Cambridge, MA 02138
(617) 495-9454
Fax: (617) 496-5747
E-mail: dani_rodrik@harvard.edu
http://www.ksg.harvard.edu/rodrik/

This version
August 2004

ABSTRACT

This is an attempt to derive broad, strategic lessons from the diverse experience with economic growth in last fifty years. The paper revolves around two key arguments. One is that neoclassical economic analysis is a lot more flexible than its practitioners in the policy domain have generally given it credit. In particular, first-order economic principles—protection of property rights, market-based competition, appropriate incentives, sound money, and so on—do not map into unique policy packages. Reformers have substantial room for creatively packaging these principles into institutional designs that are sensitive to local opportunities and constraints. Successful countries are those that have used this room wisely. The second argument is that igniting economic growth and sustaining it are somewhat different enterprises. The former generally requires a limited range of (often unconventional) reforms that need not overly tax the institutional capacity of the economy. The latter challenge is in many ways harder, as it requires constructing over the longer term a sound institutional underpinning to endow the economy with resilience to shocks and maintain productive dynamism. Ignoring the distinction between these two tasks leaves reformers saddled with impossibly ambitious, undifferentiated, and impractical policy agendas.

GROWTH STRATEGIES

Dani Rodrik

"[A]s far as the LDCs are concerned, it is probably fair to say that at least a crude sort of 'justice' prevails in the economic policy realm. Countries that have run their economies following the policy tenets of the professionals have on the whole reaped good fruit from the effort; likewise, those that have flown in the face of these tenets have had to pay the price."

-- Arnold C. Harberger (1985, p. 42)

"When you get right down to business, there aren't too many policies that we can say with certainty deeply and positively affect growth."

-- Arnold C. Harberger (2003, p. 215)

I. Introduction

Real per-capita income in the developing world grew at an average rate of 2.3 percent per annum during the four decades between 1960 and 2000.[1] This is a high growth rate by almost any standard. At this pace incomes double every 30 years, allowing each generation to enjoy a level of living standards that is twice as high as the previous generation's. To provide some historical perspective on this performance, it is worth noting that Britain's per-capita GDP grew at a mere 1.3 percent per annum during its period of economic supremacy in the middle of the 19th century (1820-1870) and that the United States grew at only 1.8 percent during the half century before World War I when it overtook Britain as the world's economic leader (Maddison 2001, Table B-22, 265). Moreover, with few exceptions, economic growth in the last few decades has been accompanied by significant improvements in social indicators such as literacy, infant mortality, life expectation, and the like. So on balance the recent growth record looks quite impressive.

However, since the rich countries themselves grew at a very rapid clip of 2.7 percent during the period 1960-2000, few developing countries consistently managed to

[1] This figure refers to the exponential growth rate of GDP per capita (in constant 1995 US$) for the group of low- and middle-income countries. The data come from the World Development Indicators 2002 CD-ROM of the Word Bank.

close the economic gap between them and the advanced nations. As Figure 1 indicates, the countries of East and Southeast Asia constitute the sole exception. Excluding China, this region experienced per-capita GDP growth of 4.4 percent over 1960-2000. Despite the Asian financial crisis of 1997-98 (which shows as a slight dip in Figure 1), countries such as South Korea, Thailand and Malaysia ended the century with productivity levels that stood significantly closer to those enjoyed in the advanced countries.

<Figure 1 here>

Elsewhere, the pattern of economic performance has varied greatly across different time periods. China has been a major success story since the late 1970s, experiencing a stupendous growth rate of 8.0 percent (as compared to 2.0 percent in 1960-80). Less spectacularly, India has roughly doubled its growth rate since the early 1980s, pulling South Asia's growth rate up to 3.3 percent in 1980-2000 from 1.2 percent in 1960-1980. The experience in other parts of the world was the mirror image of these Asian growth take-offs. Latin America and Sub-Saharan Africa both experienced robust economic growth prior to the late 1970s and early 1980s—2.9 percent and 2.3 percent respectively—but then lost ground subsequently in dramatic fashion. Latin America's growth rate collapsed in the "lost decade" of the 1980s, and has remained anemic despite some recovery in the 1990s. Africa's economic decline, which began in the second half of the 1970s, continued throughout much of the 1990s and has been aggravated by the onset of HIV/AIDS and other public-health challenges. Measures of total factor productivity run parallel to these trends in per-capita output (see Table 1).

Hence the aggregate picture hides tremendous variety in growth performance, both geographically and temporally. We have high growth countries and low growth countries; countries that have grown rapidly throughout, and countries that have experienced growth spurts for a decade or two; countries that took off around 1980 and countries whose growth collapsed around 1980.

This paper is devoted to the question: what do we learn about *growth strategies* from this rich and diverse experience? By "growth strategies" I refer to economic policies and institutional arrangements aimed at achieving economic convergence with the living standards prevailing in advanced countries. My emphasis will be less on the relationship between specific policies and economic growth—the stock-in-trade of cross-national growth empirics—and more on developing a broad understanding of the contours of successful strategies. Hence my account harks back to an earlier generation of studies that distilled operational lessons from the observed growth experience, such as Albert Hirschman's *The Strategy of Economic Development* (1958), Alexander Gerschenkron's *Economic Backwardness in Historical Perspective* (1962) or Walt Rostow's *The Stages of Economic Growth* (1965). This paper follows an unashamedly inductive approach in this tradition.

A key theme in these works, as well as in the present paper, is that growth-promoting policies tend to be context specific. We are able to make only a limited number of generalizations on the effects on growth, say, of liberalizing the trade regime,

opening up the financial system, or building more schools. The experience of the last two decades has frustrated the expectations of policy advisers who thought we had a good fix on the policies that promote growth—see the shift in mood that is reflected in the two quotes from Harberger that open this paper. And despite a voluminous literature, cross-national growth regressions ultimately do not provide us with much reliable and unambiguous evidence on such operational matters.[2] An alternative approach, and the one I adopt here, is to shift our focus to a higher level of generality and to examine the broad design principles of successful growth strategies. This entails zooming away from the individual building blocks and concentrating on how they are put together.

The paper revolves around two key arguments. One is that neoclassical economic analysis is a lot more flexible than its practitioners in the policy domain have generally given it credit. In particular, first-order economic principles—protection of property rights, contract enforcement, market-based competition, appropriate incentives, sound money, debt sustainability—do not map into unique policy packages. Good institutions are those that deliver these first-order principles effectively. There is no unique correspondence between the functions that good institutions perform and the form that such institutions take. Reformers have substantial room for creatively packaging these principles into institutional designs that are sensitive to local constraints and take advantage of local opportunities. Successful countries are those that have used this room wisely.

The second argument is that igniting economic growth and sustaining it are somewhat different enterprises. The former generally requires a limited range of (often unconventional) reforms that need not overly tax the institutional capacity of the economy. The latter challenge is in many ways harder, as it requires constructing a sound institutional underpinning to maintain productive dynamism and endow the economy with resilience to shocks over the longer term. Ignoring the distinction between these two tasks leaves reformers saddled with impossibly ambitious, undifferentiated, and impractical policy agendas.

The plan for the paper is as follows. The next section sets the stage by evaluating the standard recipes for economic growth in light of recent economic performance. Section III develops the argument that sound economic principles do not map into unique institutional arrangements and reform strategies. Section IV re-interprets recent growth experience using the conceptual framework of the previous section. Section V discusses a two-pronged growth strategy that differentiates between the challenges of igniting growth and the challenges of sustaining it. Concluding remarks are presented in section VI.

II. What we know that (possibly) ain't so

[2] Easterly (2003) provides a good overview of these studies. See also Temple (1999), Brock and Durlauf (2001), and Rodriguez and Rodrik (2001).

Development policy has always been subject to fads and fashions. During the 1950s and 1960s, "big push," planning, and import-substitution were the rallying cries of economic reformers in poor nations. These ideas lost ground during the 1970s to more market-oriented views that emphasized the role of the price system and outward-orientation.[3] By the late 1980s a remarkable convergence of views had developed around a set of policy principles that John Williamson (1990) infelicitously termed the "Washington Consensus." These principles remain at the heart of today's conventional understanding of a desirable policy framework for economic growth, even though they have been greatly embellished and expanded in the years since.

The left panel in Table 2 shows Williamson's original list, which focused on fiscal discipline, "competitive" currencies, trade and financial liberalization, privatization and deregulation. These were perceived to be the key elements of what Krugman (1995, 29) has called the "Victorian virtue in economic policy," namely "free markets and sound money". Towards the end of the 1990s, this list was augmented in the thinking of multilateral agencies and policy economists with a series of so-called second-generation reforms that were more institutional in nature and targeted at problems of "good governance." A complete inventory of these Washington Consensus-plus reforms would take too much space, and in any case the precise listing differs from source to source.[4] I have shown a representative sample of ten items (to preserve the symmetry with the original Washington Consensus) in the right panel of Table 2. They range from anti-corruption and corporate governance to social safety nets and targeted anti-poverty programs.

The perceived need for second-generation reforms arose from a combination of sources. First, there was growing recognition that market-oriented policies may be inadequate without more serious institutional transformation, in areas ranging from the bureaucracy to labor markets. For example, trade liberalization may not reallocate an economy's resources appropriately if the labor markets are "rigid" or insufficiently "flexible." Second, there was a concern that financial liberalization may lead to crises and excessive volatility in the absence of a more carefully delineated macroeconomic framework and improved prudential regulation. Hence the focus on non-intermediate exchange-rate regimes, central bank independence, and adherence to international financial codes and standards. Finally, in response to the complaint that the Washington Consensus represented a trickle-down approach to poverty, the policy framework was augmented with social policies and anti-poverty programs.

It is probably fair to say that a listing along the lines of Table 2 captures in broad brushstrokes mainstream thinking about the key elements of a growth program *circa* 2000. How does such a list fare when held against the light of contemporary growth

[3] Easterly (2001) provides an insightful and entertaining account of the evolution of thinking on economic development. See also Lindauer and Pritchett (2002) and Krueger (1997).

[4] For diverse perspectives on what the list should contain, see Stiglitz (1998), World Bank (1998), Naim (1999), Birdsall and de la Torre (2001), Kaufmann (2002), Ocampo (2002), and Kuczynski and Williamson (2003).

experience? Imagine that we gave Table 2 to an intelligent Martian and asked him to match the growth record displayed in Figure 1 and Table 1 with the expectations that the list generates. How successful would he be in identifying which of the regions adopted the standard policy agenda and which did not?

Consider first the high performing East Asian countries. Since this region is the only one that has done consistently well since the early 1960s, the Martian would reasonably guess that there is a high degree of correspondence between its policies and the list in Table 2. But he would be at best half-right. South Korea's and Taiwan's growth policies, to take two important illustrations, exhibit significant departures from the Washington Consensus. Neither country undertook significant deregulation or liberalization of their trade and financial systems well into the 1980s. Far from privatizing, they both relied heavily on public enterprises. South Korea did not even welcome direct foreign investment. And both countries deployed an extensive set of industrial policies that took the form of directed credit, trade protection, export subsidization, tax incentives, and other non-uniform interventions. Using the minimal scorecard of the original Washington Consensus (left panel of Table 2), the Martian would award South Korea a grade of 5 (out of 10) and Taiwan perhaps a 6 (Rodrik 1996).

The gap between the East Asian "model" and the more demanding institutional requirements shown on the right panel of Table 2 is, if anything, even larger. I provide a schematic comparison between the standard "ideal" and the East Asian reality in Table 3 for a number of different institutional domains such as corporate governance, financial markets, business-government relationships, and public ownership. Looking at this, the Martian might well conclude that South Korea, Taiwan, and (before them) Japan stood little chance to develop. Indeed, such were the East Asian anomalies that when the Asian financial crisis of 1997-98 struck, many observers attributed the crisis to the moral hazard, "cronyism," and other problems created by East Asian-style institutions (see MacLean 1999, Frankel 2000).

The Martian would also be led astray by China's boom since the late 1970s and by India's less phenomenal, but still significant growth pickup since the early 1980s. While both of these countries have transformed their attitudes towards markets and private enterprise during this period, their policy frameworks bear very little resemblance to what is described in Table 2. India deregulated its policy regime slowly and undertook very little privatization. Its trade regime remained heavily restricted late into the 1990s. China did not even adopt a private property rights regime and it merely appended a market system to the scaffolding of a planned economy (as discussed further below). It is hardly an exaggeration to say that had the Chinese economy stagnated in the last couple of decades, the Martian would be in a better position to rationalize it using the policy guidance provided in Table 2 than he is to explain China's actual performance.[5]

[5] Vietnam, a less well known case than China, has many of the same characteristics: rapid growth since the late 1980s as a result of heterodox reform. Vietnam has benefited from a gradual turn toward markets and greater reliance on private entrepreneurship, but as Van Arkadie and Mallon (2003) argue, it is hard to square the extensive role of the state and the nature of the property rights regime with the tenets of the Washington Consensus.

The Martian would be puzzled that the region that made the most determined attempt at remaking itself in the image of Table 2, namely Latin America, has reaped so little growth benefit out of it. Countries such as Mexico, Argentina, Brazil, Colombia, Bolivia, and Peru did more liberalization, deregulation and privatization in the course of a few years than East Asian countries have done in four decades. Figure 2 shows an index of structural reform for these and other Latin American countries, taken from Lora (2001a). The index measures on a scale from 0 to 1 the extent of trade and financial liberalization, tax reform, privatization, and labor-market reform undertaken. The regional average for the index rises steadily from 0.34 in 1985 to 0.58 in 1999. Yet the striking fact from Figure 1 is that Latin America's growth rate has remained significantly below its pre-1980 level. The Martian would be at a loss to explain why growth is now lower given that the quality of Latin America's policies, as judged by the list in Table 2, has improved so much.[6] A similar puzzle, perhaps of a smaller magnitude, arises with respect to Africa, where economic decline persists despite an overall (if less marked) "improvement" in the policy environment.[7]

<Figure 2 here>

The Martian would recognize that the growth record is consistent with some of the *higher-order* economic principles that inspire the standard policy consensus. A semblance of property rights, sound money, fiscal solvency, market-oriented incentives—these are elements that are common to all successful growth strategies.[8] Where they have been lacking, economic performance has been lackluster at best. But the Martian would also have to conclude that the mapping from our more detailed policy preferences (such

---

[6] Lora (2001b) finds that structural reforms captured by this index do correlate with growth rates in the predicted manner, but that the impacts (taking the decade of the 1990s as a whole) are not that strong. Another econometric study by Loayza et al. (2002) claims that Latin America's reforms added significantly to the region's growth. However the latter paper uses outcome variables such as trade/GDP and financial depth ratios as its indicators of "policy," and therefore is unable to link economic performance directly to the reforms themselves. Lin and Liu (2003) attribute the failure of the Washington Consensus to the non-viability of enterprises created under the previous "distorted" policy regime and the political impossibility of letting these go bust.

[7] See also Milanovic (2003) for a closely related Martian thought experiment. Milanovic emphasizes that economic growth has declined in most countries despite greater globalization.

[8] Here is how Larry Summers (2003) summarizes the recent growth evidence: "[The] rate at which countries grow is substantially determined by three things: their ability to integrate with the global economy through trade and investment; their capacity to maintain sustainable government finances and sound money; and their ability to put in place an institutional environment in which contracts can be enforced and property rights can be established. I would challenge anyone to identify a country that has done all three of these things and has not grown at a substantial rate." Note how these recommendations are couched not in terms of specific policies (maintain tariffs below *x* percent, raise the government primary surplus above *y* percent, privatize state enterprises, and so on), but in terms of "abilities" and "capacities" to get certain outcomes accomplished. I will suggest below that these "abilities" and "capacities" do not map neatly into the standard policy preferences, and can be generated in a variety of ways.

as those in Table 2) to economic success is quite imperfect. He would wonder if we cannot do better.

III. The indeterminate mapping from economic principles to institutional arrangements

Here is another thought experiment. Imagine a Western economist was invited to Beijing in 1978 in order to advise the Chinese leadership on a reform strategy. What would she recommend and why?

The economist would recognize that reform must start in the rural areas since the vast majority of the poor live there. An immediate recommendation would be the *liberalization of agricultural markets* and the *abolition of the state order system* under which peasants had to make obligatory deliveries of crops at low, state-controlled prices. But since price liberalization alone would be inadequate to generate the appropriate supply incentives under a system of communal land ownership, the economist would also recommend the *privatization of land.* Next, the economist would have to turn her attention to the broader implications of price liberalization in agriculture. Without access to cheap grains, the state would be left without a source of implicit tax revenue, so *tax reform* must be on the agenda as well. And in view of the rise of food prices, there must be a way to respond to urban workers' demand for higher wages. State enterprises in urban areas must be *corporatized*, so that their managers are in a position to adjust their wages and prices appropriately.

But now there are other problems that need attention. In an essentially closed and non-competitive economy, price-setting autonomy for the state behemoths entails the exercise of monopoly power. So the economist would likely recommend *trade liberalization* in order to "import" price discipline from abroad. Openness to trade in turn calls for other complementary reforms. There must be *financial sector reform* so that financial intermediaries are able to assist domestic enterprises in the inevitable adjustments that are called forth. And of course there must be *social safety nets* in place so that those workers who are temporarily displaced have some income support during the transition.

The story can be embellished by adding other required reforms, but the message ought to be clear. By the time the Western economist is done, the reform agenda she has formulated looks very similar to the Washington Consensus (see Table 4). The economist's reasoning is utterly plausible, which underscores the point that the Washington Consensus is far from silly: it is the result of systematic thinking about the multiple, often complementary reforms needed to establish property rights, put market incentives to work, and maintain macroeconomic stability. But while this particular reform program represents a logically consistent way achieving these end goals, it is not the only one that has the potential of doing so. In fact, in view of the administrative and political constraints that such an ambitious agenda is likely to encounter, it is not implausible that there would be better ways of getting there.

How can we be sure of this? We know this because China took a very different approach to reform—one that was experimental in nature and relied on a series of institutional innovations that departed significantly from Western norms. What is important to realize about these innovations is that in the end they delivered—for a period of a couple of decades at least—the very same goals that the Western economist would have been hoping for: market-oriented incentives, property rights, macroeconomic stability. But they did so in a peculiar fashion that, given the Chinese historical and political context, had numerous advantages.

For example, the Chinese authorities liberalized agriculture only *at the margin* while keeping the plan system intact. Farmers were allowed to sell surplus crops freely at a market-determined price only after they had fulfilled their obligations to the state under the state order system. As Lau, Qian, and Roland (2000) explain, this was an ingenious system that generated efficiency without creating any losers. In particular, it was a shortcut that neatly solved a conundrum inherent in wholesale liberalization: how to provide microeconomic incentives to producers while insulating the central government from the fiscal consequences of liberalization. As long as state quotas were set below the fully liberalized market outcome (so that transactions were conducted at market prices at the margin) and were not ratcheted up (so that producers did not have to worry about the quotas creeping up as a result of marketed surplus), China's dual-track reform in effect achieved full allocative efficiency. But it entailed a different infra-marginal distribution—one that preserved the income streams of initial claimants. The dual track approach was eventually employed in other areas as well, such as industrial goods (e.g. coal and steel) and labor markets (employment contracts). Lau et al. (2000) argue that the system was critical to achieve political support for the reform process, maintain its momentum, and minimize adverse social implications.

Another important illustration comes from the area of property rights. Rather than privatize land and industrial assets, the Chinese government implemented novel institutional arrangements such as the Household Responsibility System (under which land was "assigned" to individual households according to their size) and Township and Village Enterprises (TVEs). The TVEs were the growth engine of China until the mid-1990s (Qian 2003), with their share in industrial value added rising to more than 50 percent by the early 1990s (Lin et al. 1996, 180), so they deserve special comment. Formal ownership rights in TVEs were vested not in private hands or in the central government, but in local communities (townships or villages). Local governments were keen to ensure the prosperity of these enterprises as their equity stake generated revenues directly for them. Qian (2003) argues that in the environment characteristic of China, property rights were effectively more secure under direct local government ownership than they would have been under a private property-rights legal regime. The efficiency loss incurred due to the absence of private control rights was probably outweighed by the implicit security guaranteed by local government control. It is difficult to explain otherwise the remarkable boom in investment and entrepreneurship generated by such enterprises.

Qian (2003) discusses other examples of "transitional institutions" China employed to fuel economic growth—fiscal contracts between central and local governments, anonymous banking—and one may expand his list by including arrangements such as Special Economic Zones. The main points to take from this experience are the following. First, China relied on highly unusual, non-standard institutions. Second, these unorthodox institutions worked precisely because they produced orthodox results, namely market-oriented incentives, property rights, macroeconomic stability, and so on. Third, it is hard to argue, in view of China's stupendous growth, that a more standard, "best-practice" set of institutional arrangements would have necessarily done better.

The Chinese experience helps lay out the issues clearly because its institutional innovations and growth performance are both so stark. But China's experience with non-standard growth policies is hardly unusual; in fact it is more the rule than the exception. The (other) East Asian anomalies noted previously (Table 3) can be viewed as part of the same pattern: non-standard practices in the service of sound economic principles. I summarize a few non-Chinese illustrations in Table 5.

Consider for example the case of financial controls. I noted earlier that few of the successful East Asian countries undertook much financial liberalization early on in their development process. Interest rates remained controlled below market-clearing levels and competitive entry (by domestic or foreign financial intermediaries) was typically blocked. It is easy to construct arguments as to why this was beneficial from an economic standpoint. Table 5 summarizes the story laid out by Hellman, Morduck, and Stiglitz (1997), who coin the term "financial restraint" for the Asian model. Where asymmetric information prevails and the level of savings is sub-optimal, Hellman et al. argue that creating a moderate amount of rents for incumbent banks can generate useful incentives. These rents induce banks to do a better job of monitoring their borrowers (since there is more at stake) and to expand effort to mobilize deposits (since there are rents to be earned on them). The quality and level of financial intermediation can both be higher than under financial liberalization. These beneficial effects are more likely to materialize when the pre-existing institutional landscape has certain properties—for example when the state is not "captured" by private interests and the external capital account is restricted (see last two columns of Table 5). When these preconditions are in place, the economic logic behind financial restraint is compelling.

The second illustration in Table 5 comes from South Korea's and Taiwan's experiences with industrial policy. The governments in these countries rejected the standard advice that they take an arms' length approach to their enterprises and actively sought to coordinate private investments in targeted sectors. Once again, it is easy to come up with economic models that provide justification for this approach. In Rodrik (1995), I argued that the joint presence of scale economies and inter-industry linkages can depress the private return to investment in non-traditional activities below the social return. Industrial policy can be viewed as a "coordination device" to stimulate socially profitable investments. In particular, the socialization of investment risk through implicit bailout guarantees may be economically beneficial despite the obvious moral hazard risk

it poses. However, once again, there are certain prerequisites and institutional complements that have to be in place for this approach to make sense (see Table 5).

The third illustration in Table 5 refers to Japan and concerns the internal organization of the workplace, drawing on Aoki's (1997) work. Aoki describes the peculiar institutional foundations of Japan's postwar success as having evolved from a set of arrangements originally designed for wartime mobilization and centralized control of resources. He presents Japan's team-centered approach to work organization and its redistribution of economic resources from advanced to backward sectors—arrangements that he terms "horizontal hierarchy" and "bureau-pluralism," respectively—as solutions to particular informational and distributive dilemmas the Japanese economy faced in the aftermath of World War II. Unlike the previous authors, however, he views this fit between institutions and economic challenges as having been unintended and serendipitous.

Lest the reader think this is solely an East Asian phenomenon, an interesting example of institutional innovation comes from Mauritius (Rodrik 1999). Mauritius owes a large part of its success to the creation in 1970 of an export-processing zone (EPZ), which enabled an export boom in garments to European markets. Yet, instead of liberalizing its trade regime across the board, Mauritius combined this EPZ with a domestic sector that was highly protected until the mid-1980s, a legacy of the policies of import-substituting industrialization (ISI) followed during the 1960s. The industrialist class that had been created with these policies was naturally opposed to the opening up of the trade regime. The EPZ scheme provided a neat way around this difficulty (Wellisz and Saw 1993). The creation of the EPZ generated new profit opportunities, without taking protection away from the import-substituting groups. The segmentation of labor markets was particularly crucial in this regard, as it prevented the expansion of the EPZ (which employed mainly female labor) from driving wages up in the rest of the economy, and thereby disadvantaging import-substituting industries. New profit opportunities were created at the margin, while leaving old opportunities undisturbed. At a conceptual level, the story here is essentially very similar to the two-track reforms in China described earlier. To produce the results it did, however, the EPZ also needed a source of investible funds, export-oriented expertise, and market access abroad, which were in turn provided by a terms-of-trade boom, entrepreneurs from Hong Kong, and preferential market access in Europe, respectively (Rodrik 1999; Subramanian and Roy 2003).

In reviewing cases such as these, there is always the danger of reading too much into them after the fact. In particular, we need to avoid several fallacies. First, we cannot simply assume that institutions take the form that they do <u>because</u> of the functions that they perform (the functionalist fallacy). Aoki's account of Japan is a particularly useful reminder that a good fit between form and function might be the unintended consequence of historical forces. Second, it is not correct to ascribe the positive outcomes in the cases just reviewed only to their anomalies (the ex-post rationalization fallacy). Many accounts of East Asian success emphasize the standard elements--fiscal conservatism, investment in human resources, and export orientation (see for example World Bank 1993). As I will discuss below, East Asian institutional anomalies have often produced perverse

results when employed in other settings. And it is surely not the case that all anomalies are economically functional.

The main point I take from these illustrations is robust to these fallacies, and has to do with the "plasticity" of the institutional structure that neoclassical economics is capable of supporting. All of the above institutional anomalies are compatible with, and can be understood in terms of, neoclassical economic reasoning ("good economics"). Neoclassical economic analysis does not determine the form that institutional arrangements should or do take. What China's case and other examples discussed above demonstrate is that the higher-order principles of sound economic management do not map into unique institutional arrangements.

In fact, principles such as appropriate incentives, property rights, sound money, and fiscal solvency all come institution-free. We need to operationalize them through a set of policy actions. The experiences above show us that there may be multiple ways of packing these principles into institutional arrangements. Different packages have different costs and benefits depending on prevailing political constraints, levels of administrative competence, and market failures. The pre-existing institutional landscape will typically offer both constraints and opportunities, requiring creative shortcuts or bold experiments. From this perspective, the "art" of reform consists of selecting appropriately from a potentially infinite menu of institutional designs.

A direct corollary of this line of argument is that there is only a weak correspondence between the higher-order principles of neoclassical economics and the specific policy recommendations in the standard list (as enumerated in Table 2). To see this, consider for example one of the least contentious recommendations in the list, having to do with trade liberalization. Can the statement "trade liberalization is good for economic performance" be derived from first principles of neoclassical economics? Yes, but only if a number of side conditions are met:

- The liberalization must be complete or else the reduction in import restrictions must take into account the potentially quite complicated structure of substitutability and complementarity across restricted commodities.[9]
- There must be no microeconomic market imperfections other than the trade restrictions in question, or if there are some, the second-best interactions that are entailed must not be adverse.[10]

[9] There is a large theoretical literature on partial trade reform, which shows the difficulty of obtaining unambiguous characterizations of the welfare effects of incomplete liberalization. See Hatta (1977), Anderson and Neary (1992), and Lopez and Panagariya (1993). For an applied general equilibrium analysis of how these issues can complicate trade reform in practice, see Harrison, Rutherford, and Tarr (1993).

[10] For an interesting empirical illustration on how trade liberalization can interact adversely with environmental externalities, see Lopez (1997).

- The home economy must be "small" in world markets, or else the liberalization must not put the economy on the wrong side of the "optimum tariff."[11]
- The economy must be in reasonably full employment, or if not, the monetary and fiscal authorities must have effective tools of demand management at their disposal.
- The income redistributive effects of the liberalization should not be judged undesirable by society at large, or if they are, there must be compensatory tax-transfer schemes with low enough excess burden.[12]
- There must be no adverse effects on the fiscal balance, or if there are, there must be alternative and expedient ways of making up for the lost fiscal revenues.
- The liberalization must be politically sustainable and hence credible so that economic agents do not fear or anticipate a reversal.[13]

All these theoretical complications could be sidestepped if there were convincing evidence that in practice trade liberalization systematically produces improved economic performance. But even for this relatively uncontroversial policy, it has proved difficult to generate unambiguous evidence (see Rodriguez and Rodrik 2001, Vamvakidis 2002, and Yanikkaya 2003).[14]

The point is that even the simplest of policy recommendations—"liberalize foreign trade"—is contingent on a large number of judgment calls about the economic and political context in which it is to be implemented.[15] Such judgment calls are often made implicitly. Rendering them explicit has a double advantage: it warns us about the potential minefields that await the standard recommendations, and it stimulates creative

[11] This is not a theoretical curiosum. Gilbert and Varangis (2003) argue that the liberalization of cocoa exports in West African countries has depressed world cocoa prices, with most of the benefits being captured by consumers in developed countries.

[12] The standard workhorse model of international trade, the factor-endowments model and its associated Stolper-Samuelson theorem, comes with sharp predictions on the distributional effects of import liberalization (the "magnification effect").

[13] Calvo (1989) was the first to point out that lack of credibility acts as an intertemporal distortion. See also Rodrik (1991).

[14] Recent empirical studies have begun to look for non-linear effects of trade liberalization. In a study of India's liberalization, Aghion et al. (2003) find that trade liberalization appears to have generated differentiated effects across Indian firms depending on prevailing industrial capabilities and labor market regulations. Firms that were close to the technological frontier and in states with more "flexible" regulations responded positively while others responded negatively. See also Helleiner (1994) for a useful collection of country studies that underscores the contingent nature of economies' response to trade liberalization.

[15] This is one reason why policy discussions on standard recommendations such as trade liberalization and privatization now often take the formulaic form: "policy *x* is not a panacea; in order to work, it must be supported by reforms in the areas of *a, b, c, d,* and so on."

thinking on alternatives (as in China) that can sidestep those minefields. By contrast, when the policy recommendation is made unconditionally, as in the Washington Consensus, the gamble is that the policy's prerequisites will coincide with our actual draw from a potentially large universe of possible states of the world.

I summarize this discussion with the help of Tables 6, 7, and 8 dealing with microeconomic policy, macroeconomic policy, and social policy, respectively. Each table contains three columns. The first column displays the ultimate goal that is targeted by the policies and institutional arrangements in the three domains. Hence microeconomic policies aim to achieve static and dynamic efficiency in the allocation of resources. Macroeconomic policies aim for macroeconomic and financial stability. Social policies target poverty reduction and social protection.

The next column displays some of the key higher-order principles that economic analysis brings to the table. Allocative efficiency require property rights, the rule of law, and appropriate incentives. Macroeconomic and financial stability requires sound money, fiscal solvency, and prudential regulation. Social inclusion requires incentive compatibility and appropriate targeting. These are the "universal principles" of sound economic management. They are universal in the sense that it is hard to see what any country would gain by systematically defying them. Countries that have adhered to these principles—no matter how unorthodox their manner of doing so may have been—have done well while countries that have flouted them have typically done poorly.

From the standpoint of policy makers, the trouble is that these universal principles are not operational as stated. In effect, the answers to the real questions that preoccupy policy makers—how far should I go in opening up my economy to foreign competition, should I free up interest rates, should I rely on payroll taxes or the VAT, and the others listed in the third column of each table--cannot be directly deduced from these principles. This opens up space for a multiplicity of institutional arrangements that are compatible with the universal, higher-order principles.

These tables clarify why the standard recommendations (Table 2) correlates poorly with economic performance around the world. The Washington Consensus, in its various forms, has tended to blur the line that separates column 2 from column 3. Policy advisors have been too quick in jumping from the higher-order principles in column 2 to taking unconditional stands on the specific operational questions posed in column 3. And as their policy advice has yielded disappointing results, they have moved on to recommendations with even greater institutional specificity (as with "second generation reforms"). As a result, sound economics has often been delivered in unsound form.

I emphasize that this argument is not one about the advantages of gradualism over shock therapy. In fact, the set of ideas I have presented are largely orthogonal to the long-standing debate between the adherents of the two camps (see for example Lipton and Sachs 1990, Aslund et al. 1996, Williamson and Zagha 2002). The strategy of gradualism presumes that policy makers have a fairly good idea of the institutional arrangements that they want to acquire ultimately, but that for political and other reasons

they can proceed only step-by-step in that direction. The argument here is that there is typically a large amount of uncertainty about what those institutional arrangements are, and therefore that the process that is required is more one of "search and discovery" than one of gradualism. The two strategies may coincide when policy changes reveal information and small-scale policy reforms have a more favorable ratio of information revelation to risk of failure.[16] But it is best not to confuse the two strategies. What stands out in the real success cases, as I will further illustrate below, is not gradualism per se but an unconventional mix of standard and non-standard policies well attuned to the reality on the ground.

IV. Back to the real world

Previously we had asked our Martian to interpret economic performance in the real world from the lens of the standard reform agenda. Suppose we now remove the constraint and ask him to summarize the stylized facts as he sees them. Here is a list of four stylized facts that he may come up with.

1. In practice, growth spurts are associated with a narrow range of policy reforms.

One of the most encouraging aspects of the comparative evidence on economic growth is that it often takes very little to get growth started. To appreciate the point, it is enough to turn to Table 9, which lists 83 cases of growth accelerations. The table shows all cases of significant growth accelerations since the mid-1950s that can be identified statistically. The definition of a growth acceleration is the following: an increase in an economy's per-capita GDP growth of 2 percentage points or more (relative to the previous 5 years) that is sustained over at least 8 years. The timing of the growth acceleration is determined by fitting a spline centered on the candidate break years, and selecting the break that maximizes the fit of the equation (see Hausmann, Pritchett, and Rodrik 2004 for details on the procedure).[17]

Most of the usual suspects are included in the table: for example Taiwan 1961, Korea 1962, Indonesia 1967, Brazil 1967, Mauritius 1971, China 1978, Chile 1986, Uganda 1989, Argentina 1990, and so on. But the exercise also yields a large number of much less well-known cases, such as Egypt 1976 or Pakistan 1979. In fact, the large number of countries that have managed to engineer at least one instance of transition to high growth may appear as surprising. As I will discuss later, most of these growth spurts have eventually collapsed. Nonetheless, an increase in growth of 2 percent (and typically more) over the better part of a decade is nothing to sneer at, and it is worth asking what produces it.

---

[16] For example, Dewatripont and Roland (1995) and Wei (1997) present models in which gradual reforms reveal information and affect subsequent political constraints.

[17] The selection strategy allows multiple accelerations, but they must be at least five years apart. We require post-acceleration growth to be at least 3.5 percent, and also rule out recoveries from crises.

In the vast majority of the cases listed in Table 9, the "shocks" (policy or otherwise) that produced the growth spurts were apparently quite mild. Asking most development economists about the policy reforms of Pakistan in 1979 or Syria in 1969 would draw a blank stare. This reflects the fact that not much reform was actually taking pace in these cases. Relatively small changes in the background environment can yield significant increase in economic activity.

Even in the well-known cases, policy changes at the outset have been typically modest. The gradual, experimental steps towards liberalization that China undertook in the late 1970s were discussed above. South Korea's experience in the early 1960s was similar. The military government led by Park Chung Hee that took power in 1961 did not have strong views on economic reform, except that it regarded economic development as its key priority. It moved in a trial-and-error fashion, experimenting at first with various public investment projects. The hallmark reforms associated with the Korean miracle, the devaluation of the currency and the rise in interest rates, came in 1964 and fell far short of full liberalization of currency and financial markets. As these instances illustrate, an attitudinal change on the part of the top political leadership towards a more market-oriented, private-sector-friendly policy framework often plays as large a role as the scope of policy reform itself (if not larger). Perhaps the most important example of this can be found in India: such an attitudinal change appears to have had a particularly important effect in the Indian take-off of the early 1980s, which took place a full decade before the liberalization of 1991 (de Long 2003; Rodrik and Subramanian 2004).

This is good news because it suggests countries do not need an extensive set of institutional reforms in order to start growing. Instigating growth is a lot easier in practice than the standard recipe, with its long list of action items, would lead us to believe. This should not be surprising from a growth theory standpoint. When a country is so far below its potential steady-state level of income, even moderate movements in the right direction can produce a big growth payoff. Nothing could be more encouraging to policy makers, who are often overwhelmed and paralyzed by the apparent need to undertake policy reforms on a wide and ever-expanding front.

2. The policy reforms that are associated with these growth transitions typically combine elements of orthodoxy with unorthodox institutional practices.

No country has experienced rapid growth without minimal adherence to what I have termed higher-order principles of sound economic governance—property rights, market-oriented incentives, sound money, fiscal solvency. But as I have already argued, these principles were often implemented via policy arrangements that are quite unconventional. I illustrated this using examples such as China's two-track reform strategy, Mauritius' export processing zone, and South Korea's system of "financial restraint."

It is easy to multiply the examples. When Taiwan and South Korea decided to reform their trade regimes to reduce anti-export bias, they did this not via import

liberalization (which would have been a Western economist's advice) but through selective subsidization of exports. When Singapore decided to make itself more attractive to foreign investment, it did this not by reducing state intervention but by greatly expanding public investment in the economy and through generous tax incentives (Young 1992). Botswana, which has an admirable record with respect to macroeconomic stability and the management of its diamond wealth, also has one of the largest levels of government spending (in relation to GDP) in the world. Chile, a country that is often cited as a paragon of virtue by the standard check list, has also departed from it in some important ways: it has kept its largest export industry (copper) under state ownership; it has maintained capital controls on financial inflows through the 1990s; and it has provided significant technological, organizational, and marketing assistance to its fledgling agro-industries.

In all these instances, standard desiderata such as market liberalization and outward orientation were combined with public intervention and selectivity of some sort. The former element in the mix ensures that any economist so inclined can walk away from the success cases with a renewed sense that the standard policy recommendations really "work." Most egregiously, China's success is often attributed to its turn towards market—which is largely correct—and then with an unjustified leap of logic is taken as a vindication of the standard recipe—which is largely incorrect. It is not clear how helpful such evaluations are when so much of what these countries did is unconventional and fits poorly with the standard agenda.[18]

It is difficult to identify cases of high growth where unorthodox elements have not played a role. Hong Kong is probably the only clear-cut case. Hong Kong's government has had a hands-off attitude towards the economy in almost all areas, the housing market being a major exception. Unlike Singapore, which followed a free trade policy but otherwise undertook extensive industrial policies, Hong Kong's policies have been as close to laissez-faire as we have ever observed. However, there were important prerequisites to Hong Kong's success, which illuminate once again the context-specificity of growth strategies. Most important, Hong Kong's important entrepôt role in trade, the strong institutions imparted by the British, and the capital flight from communist China had already transformed the city-state into a high investment, high entrepreneurship economy by the late 1950s. As Figure 3 shows, during the early 1960s Hong Kong's investment rate was more than three times higher than that in South Korea or Taiwan. The latter two economies would not reach Hong Kong's 1960 per-capita GDP until the early 1970s.[19] Hence Hong Kong did not face the same challenge that Taiwan, South Korea, and Singapore did to crowd in private investment and stimulate entrepreneurship.

---

[18] Another source of confusion is the mixing up of policies with outcomes. Successful countries end up with much greater participation in the world economy, thriving private sectors, and a lot of financial intermediation. What we need to figure out, however, are the policies that produce these results. It would be a great distortion of the strategy followed by countries such as China, South Korea, Taiwan and others to argue that these outcomes were the result of trade and financial liberalization, and privatization.

[19] These and investment data are from the Penn World Tables 6.1.

<Figure 3 here>

It goes without saying that not all unorthodox remedies work. And those that work sometimes do so only for a short while. Consider for example Argentina's experiment in the 1990s with a currency board. Most economists would consider a currency board regime as too risky for an economy of Argentina's size insofar as it prevents expenditure switching via the exchange rate. (Hong Kong has long operated a successful marketing board.) However, as the Argentinean economy began to grow rapidly in the first half of the 1990s, many analysts altered their views. Had the Asian crisis of 1997-98 and the Brazilian devaluation of 1999 not forced Argentina off its currency board, it would have been easy to construct a story ex post about the virtues of the currency board as a growth strategy. The currency board sought to counteract the effects of more than a century of financial mismanagement through monetary discipline. It was a shortcut aimed at convincing foreign and domestic investors that the rules of the game had changed irrevocably. Under better external circumstances, the credibility gained might have more than offset the disadvantages. The problem in this case was the unwillingness to pull back from the experiment even when it became clear that the regime had left the Argentine economy with a hopelessly uncompetitive real exchange rate. The lesson is that institutional innovation requires a pragmatic approach which avoids ideological lock-in.

3. Institutional innovations do not travel well.

The more discouraging aspect of the stylized facts is that the policy packages associated with growth accelerations—and particularly the elements therein that are non-standard—tend to vary considerably from country to country. China's two-track strategy of reform differs significantly from India's gradualism. South Korea's and Taiwan's more protectionist trade strategy differs markedly from the open trade policies of Singapore (and Hong Kong). Even within strategies that look superficially similar, closer look reveals large variation. Taiwan and South Korea both subsidized non-traditional industrial activities, but the former did it largely through tax incentives and the latter largely through directed credit.[20]

Attempts to emulate successful policies elsewhere often fail. When Gorbachev tried to institute a system similar to China's Household Responsibility System and two-track pricing in the Soviet Union during the mid- to late-1980s, it produced few of the beneficial results that China had obtained.[21] Most developing countries have export processing zones of one kind or another, but few have been as successful as the one in

[20] On the institutional differences among East Asian economies, see Haggard (2003).

[21] Murphy, Shleifer, and Vishny (1992) analyze this failure and attribute it to the inability of the Soviet state to enforce the plan quotas once market pricing was allowed (albeit at the margin). This had been critical to the success of the Chinese approach.

Mauritius. Import-substituting industrialization (ISI) worked in Brazil, but not in Argentina.[22]

In light of the arguments made earlier, this experience should not be altogether surprising. Successful reforms are those that package sound economic principles around local capabilities, constraints and opportunities. Since these local circumstances vary, so do the reforms that work. An immediate implication is that growth strategies require considerable local knowledge. It does not take a whole lot of reform to stimulate economic growth—that is the good news. The bad news is that it may be quite difficult to identify where the binding constraints or promising opportunities lie. A certain amount of policy experimentation may be required in order to discover what will work. China represents the apotheosis of this experimental approach to reform. But it is worth noting that many other instances of successful reform were preceded by failed experiments. In South Korea, President Park's developmental efforts initially focused on the creation of white elephant industrial projects that ultimately went nowhere (Soon 1994, 27-28). In Chile, Pinochet's entire first decade can be viewed as a failed experiment in "global monetarism."

Economists can have a useful role to play in this process: they can identify the sources of inefficiency, describe the relevant trade offs, figure out general-equilibrium implications, predict behavioral responses, and so on. But they can do these well only if their analysis is adequately embedded within the prevailing institutional and political reality. The hard work needs to be done at home.

4. Sustaining growth is more difficult than igniting it, and requires more extensive institutional reform.

The main reason that few of the growth accelerations listed in Table 9 are etched in the consciousness of development economists is that most of them did not prove durable. In fact, as discussed earlier, over the last four decades few countries except for a few East Asian ones have steadily converged to the income levels of the rich countries. The vast majority of growth spurts tend to run out of gas after a while. The experience of Latin America since the early 1980s and the even more dramatic collapse of Sub-Saharan Africa are emblematic of this phenomenon. In a well-known paper, Easterly, Kremer, Pritchett and Summers (1993) were the first to draw attention to a related finding, namely the variability in growth performance across time periods. The same point is made on a broader historical canvas by Goldstone (forthcoming).

Hence growth in the short- to medium-term does not guarantee success in the long-term. A plausible interpretation is that the initial reforms need to be deepened over time with efforts aimed at strengthening the institutional underpinning of market economies. It would be nice if a small number of policy changes—which, as argued above, is what produces growth accelerations—could produce growth over the longer

[22] TFP growth averaged 2.9 and 0.2 percent per annum in Brazil and Argentina, respectively, during 1960-73. See Rodrik (1999) and Collins and Bosworth (1996).

term as well, but this is obviously unrealistic. I will discuss some of the institutional prerequisites of sustained growth in greater detail later in the paper. But the key to longer-term prosperity, once growth is launched, is to develop institutions that maintain productive dynamism and generate resilience to external shocks.

For example, the growth collapses experienced by many developing countries in the period from the mid-1970s to the early 1980s seem to be related mainly to the inability to adjust to the volatility exhibited by the external environment at that time. In these countries, the effects of terms-of-trade and interest-rate shocks were magnified by weak institutions of conflict management (Rodrik 1999b). This, rather than the nature of microeconomic incentive regimes in place (e.g., import substituting industrialization), is what caused growth in Africa and Latin America to grind to a halt after the mid-1970s and early 1980s (respectively). The required macroeconomic policy adjustments set off distributive struggles and proved difficult to undertake. Similarly, the weakness of Indonesia's institutions explains why that country could not extricate itself from the 1997-98 East Asian financial crisis (see Temple 2003), while South Korea, for example, did a rapid turnaround. These examples are also a warning that continued growth in China cannot be taken for granted: without stronger institutions in areas ranging from financial markets to political governance, the Chinese economy may well find itself having outgrown its institutional underpinnings.[23]

V. A two-pronged growth strategy

As the evidence discussed above reveals, growth accelerations are feasible with minimal institutional change. The deeper and more extensive institutional reforms needed for long-term convergence take time to implement and mature. And they may not be the most effective way to raise growth at the outset because they do not directly target the most immediate constraints and opportunities facing an economy. At the same time, such institutional reforms can be much easier to undertake in an environment of growth rather than stagnation. These considerations suggest that successful growth strategies are based on a two-pronged effort: a short-run strategy aimed at stimulating growth, and a medium- to long-run strategy aimed at sustaining growth.[24] The rest of this section takes these up in turn.

1. An investment strategy to kick-start growth

From the standpoint of economic growth, the most important question in the short run for an economy stuck in a low-activity equilibrium is: how do you get entrepreneurs

[23] Young (2000) argues that China's reform strategy may have made things worse in the long run, by increasing the number of distorted margins.

[24] A similar distinction is also made by Ocampo (2003), who emphasizes that many of the long-run correlates of growth (such as improved institutions) are the result, and not the instigator, of growth. There is also an analogue in the political science literature in the distinction between the political prerequisites of initiating and sustaining reform (see Haggard and Kaufman 1983).

excited about investing in the home economy? "Invest" here has to be interpreted broadly, as referring to all the activities that entrepreneurs undertake, such as expanding capacity, employing new technology, producing new products, searching for new markets, and so on. As entrepreneurs become energized, capital accumulation and technological change are likely to go hand in hand—too entangled with each other to separate out cleanly.

What sets this process into motion? There are two kinds of views on this in the literature. One approach emphasizes the role of government-imposed barriers to entrepreneurship. In this view, policy biases towards large and politically-connected firms, institutional failures (in the form of licensing and other regulatory barriers, inadequate property rights and contract enforcement), and high levels of policy uncertainty and risk create dualistic economic structures and repress entrepreneurship. The removal of the most egregious forms of these impediments is then expected to unleash a flurry of new investments and entrepreneurship. According to the second view, the government has to play a more pro-active role than simply getting out of the private sector's way: it needs to find means of crowding in investment and entrepreneurship with some positive inducements. In this view, economic growth is not the natural order of things, and establishing a fair and level playing field may not be enough to spur productive dynamism. The two views differ in the importance they attach to prevailing, irremovable market imperfections and their optimism with regard to governments' ability to design and implement appropriate policy interventions.

(a) Government failures

A good example of the first view is provided by the strategy of development articulated in Stern (2001). In a deliberate evocation of Hirschman's *The Strategy of Economic Development* (1958), Stern outlines an approach with two pillars: building an appropriate "investment climate" and "empowering poor people." The former is the relevant part of his approach in this context. Stern defines "investment climate" quite broadly, as "the policy, institutional, and behavioral environment, both present and expected, that influences the returns and risks associated with investment" (2001, 144-45). At the same time, he recognizes the need for priorities and the likelihood that these priorities will be context specific. He emphasizes the favorable dynamics that are unleashed once a few, small things are done right.

In terms of actual policy content, Stern's illustrations make clear that he views the most salient features of the investment climate to be government-imposed imperfections: macroeconomic instability and high inflation, high government wages that distort the functioning of labor markets, a large tax burden, arbitrary regulations, burdensome licensing requirements, corruption, and so on. The strategy he recommends is to use enterprise surveys and other techniques to uncover which of these problems bite the most, and then to focus reforms on the corresponding margin. Similar perspectives can be found in Johnson et al. (2000), Friedman et al. (2000), and Aslund and Johnson (2003). Besley and Burgess (2002) provide evidence across Indian states on the productivity depressing effects of labor market regulations. The title of Shleifer and Vishny's (1998)

book aptly summarizes the nature of the relevant constraint in this view: The Grabbing Hand: Government Pathologies and Their Cures.

(b) Market failures

The second approach focuses not on government-imposed constraints, but on market imperfections inherent in low-income environments that block investment and entrepreneurship in non-traditional activities. In this view, economies can get stuck in a low-level equilibrium due to the nature of technology and markets, even when government policy does not penalize entrepreneurship. There are many versions of this latter approach, and some of the main arguments are summarized in the taxonomy presented in Table 10. I distinguish here between stories that are based on learning spillovers (a non-pecuniary externality) and those that are based on market-size externalities induced by scale economies. See also the useful discussion of these issues in Ocampo (2003), which takes a more overtly structuralist perspective.

As Acemoglu, Aghion, and Zilibotti (2002) point out, two types of learning are relevant to economic growth: (a) adaptation of existing technologies; and (b) innovation to create new technologies. Early in the development process, the kind of learning that matters the most is of the first type. There are a number of reasons why such learning can be subject to spillovers. There may be a threshold level of human capital beyond which the private return to acquiring skills becomes strongly positive (as in Azariadis and Drazen 1990). There may be learning-by-doing which is either external to individual firms, or cannot be properly internalized due to imperfections in the market for credit (as in Matsuyama 1992). Or there may be learning about a country's own cost structure, which spills over from the incumbents to later entrants (as in Hausmann and Rodrik 2002). In all these cases, the relevant learning is under-produced in a decentralized equilibrium, with the consequence that the economy fails to diversify into non-traditional, more advanced lines of activity.[25] There then exist policy interventions that can improve matters. With standard externalities, the first-best takes the form of a corrective subsidy targeted at the relevant distorted margin. In practice, revenue, administrative or informational constraints may make resort to second-best interventions inevitable.

For example, Hausmann and Rodrik (2002) suggest a carrot-and-stick strategy to deal with the learning barrier to industrialization that they identify. In that model, costs of production in non-traditional activities are uncertain, and they are revealed only after an upfront investment by an incumbent. Once that initial investment is made, the cost information becomes public knowledge. Entrepreneurs engaged in the cost discovery process incur private costs, but provide social benefits that can vastly exceed their anticipated profits. The first-best policy here, which is an entry subsidy, suffers from an

[25] Imbs and Wacziarg (2003) demonstrate that sectoral diversification is a robust correlate of economic growth at lower levels of income. This is in tension with standard models of trade and specialization under constant returns to scale. Sectoral concentration starts to increase only after a relatively high level of income is reached, with the turning point coming somewhere between $8,500 and $9,500 in 1985 U.S. dollars.

inextricable moral hazard problem. Subsidized entrants have little incentive to engage subsequently in costly activities to discover costs. A second-best approach takes the form of incentives contingent on good performance. Hausmann and Rodrik (2002) evaluate East Asian and Latin American industrial policies from this perspective. They argue that East Asian policies were superior in that they effectively combined incentives with discipline. The former was provided through subsidies and protection, while the latter was provided through government monitoring and the use of export performance as a productivity yardstick. Latin American firms under import substituting industrialization (ISI) received considerable incentives, but faced very little discipline. In the 1990s, these same firms arguably faced lots of discipline (exerted through foreign competition), but little incentives. This line of argument provides one potential clue to the disappointing economic performance of Latin America in the 1990s despite a much improved "investment climate" according to the standard criteria.

The second main group of stories shown in Table 10 relates to the existence of coordination failures induced by scale economies. The big-push theory of development, articulated first by Rosenstein-Rodan (1943) and formalized by Murphy, Shleifer and Vishny (1989), is based on the idea that moving out of a low-level steady state requires coordinated and simultaneous investments in a number of different areas. A general formulation of such models can provided as follows. Let the level of profits in a given modern-sector activity depend on n, the proportion of the economy that is already engaged in modern activities: $\pi^m(n)$, with $d\pi^m(n)/dn > 0$. Let profits in traditional activities be denoted π^t. Suppose modern activities are unprofitable for an individual entrant if no other entrepreneur already operates in the modern sector, but highly profitable if enough entrepreneurs do so: $\pi^m(0) < \pi^t$ and $\pi^m(1) > \pi^t$. Then $n = 0$ and $n = 1$ are both possible equilibria, and industrialization may never take hold in an economy that starts with $n = 0$. The precise mechanism that generates profit functions of this form depends on the model in question. Murphy, Shleifer, and Vishny (1989) develop models in which the complementarity arises from demand spillovers across final goods produced under scale economies or from bulky infrastructure investments. Rodriguez-Clare (1996), Rodrik (1996), and Trindade (2003) present models in which the effect operates through vertical industry relationships and specialized intermediate inputs. Hoff and Stiglitz (2001) discuss a large class of models with coordination failure characteristics.

The policy implications of such models can be quite unconventional, requiring the crowding in of private investment through subsidization, jawboning, public enterprises and the like. Despite the "big push" appellation, the requisite policies need not be wide-ranging. For example, socializing investment risk through implicit investment guarantees, a policy followed in South Korea, is welfare enhancing in Rodrik's (1996) framework because it induces simultaneous entry into the modern sector. It is also costless to the government, because the guarantees are never called on insofar as the resulting investment boom pays for itself. Hence, when successful, such policies will leave little trail on government finances or elsewhere.[26]

[26] On South Korea's implicit investment guarantees, see Amsden (1989). During the Asian financial crisis, these guarantees became an issue and they were portrayed as evidence of crony capitalism (MacLean 1999).

Both types of models listed in Table 10 suggest that the propagation of modern, non-traditional activities is not a natural process and that it may require positive inducements. One such inducement that has often worked in the past is a sizable and sustained depreciation of the real exchange rate. For a small open economy, the real exchange rate is defined as the relative price of tradables to non-tradables. In practice, this price ratio tends to move in tandem with the nominal exchange rate, the price of foreign currency in terms of home currency. Hence currency devaluations (supported by appropriate monetary and fiscal policies) increase the profitably of tradable activities across the board. From the current perspective, this has a number of distinct advantages. Most of the gains from diversification into non-traditional activities are likely to lie within manufactures and natural resource based products (i.e., tradables) rather than services and other non-tradables. Second, the magnitude of the inducement can be quite large, since sustained real depreciations of 50 percent or more are quite common. Third, since tradable activities face external competition, the activities that are encouraged tend to be precisely the ones that face the greatest market discipline. Fourth, the manner in which currency depreciation subsidizes tradable activities is completely market-friendly, requiring no micromanagement on the part of bureaucrats. For all these reasons, a credible, sustained real exchange rate depreciation may constitute the most effective industrial policy there is.

Large real exchange rate changes have played a big role in some of the more recent growth accelerations. Figure 4 shows two well-known cases: Chile and Uganda since the mid-1980s. In both cases, a substantial swing in relative prices in favor of tradables accompanied the growth take-off. In Chile, the more than doubling of the real exchange rate following the crisis of 1982-83 (the deepest in Latin America at the time) is commonly presumed to have played an instrumental role in promoting diversification into non-traditional exports and stimulating economic growth. It is worth noting that import tariffs were raised significantly as well (during 1982-85), giving import-substituting activities an additional boost. As the bottom panel of Figure 3 shows, the depreciation In Uganda was even larger. These depreciations are unlikely to have been the result of growth, since growth typically generates an appreciation of the real exchange rate through the Balassa-Samuelson effect. By contrast, large real depreciations did not play a major role in early growth accelerations in East Asia during the 1960s (Rodrik 1997).[27]

<Figure 4 here>

(c) Where to start?

The two sets of views outlined above—the government failure and market failure approaches—can help frame policy discussions and identify important ways of thinking

---

[27] Polterovich and Popov (2002) provide theory and evidence on the role of real exchange rate undervaluations in generating economic growth.

about policy priorities in the short run. The most effective point of leverage for stimulating growth obviously depends on local circumstances. It is tempting to think that the right first step is to remove government-imposed obstacles to entrepreneurial activity before worrying about "crowding in" investments through positive inducements. But this may not always be a better strategy. Certainly when inflation is in triple digits or the regulatory framework is so cumbersome that it stifles any private initiative, removing these distortions will be the most sensible initial step. But beyond that, it is difficult to say in general where the most effective margin for change lies. Asking businessmen their views on the priorities can be helpful, but not decisive. When learning spillovers and coordination failures block economic take-off, enterprise surveys are unlikely to be revealing unless the questions are very carefully crafted to elicit relevant responses.

One of the lessons of recent economic history is that creative interventions can be remarkably effective even when the "investment climate," judged by standard criteria, is pretty lousy. South Korea's early reforms took place against the background of a political leadership that was initially quite hostile to the entrepreneurial class.[28] China's TVEs have been stunningly successful despite the absence of private property rights and an effective judiciary. Conversely, the Latin American experience of the 1990s indicates that the standard criteria do not guarantee an appropriate investment climate. Governments can certainly deter entrepreneurship when they try to do too much; but they can also deter entrepreneurship when they do too little.

It is sometimes argued that heterodoxy requires greater institutional strength and therefore lies out of reach of most developing countries. But the evidence does not provide much support for this view. It is true that the selective interventions I have discussed in the case of South Korea and Taiwan were successful in part due to unusual and favorable circumstances. But elsewhere, heterodoxy served to make virtue out of institutional weakness. This is the case with China's TVEs, Mauritius' export processing zone, and India's gradualism. In these countries, it was precisely institutional weakness that rendered the standard remedies impractical. It is in part because the standard reform agenda is institutionally so highly demanding—a fact now recognized through the addition of so-called "second generation reforms"—that successful growth strategies are so often based on unconventional elements (in their early stages at least).

It is nonetheless true that the implementation of the market failure approach requires a reasonably competent and non-corrupt government. For every South Korea, there are many Zaires where policy activism is an excuse for politicians to steal and plunder. Finely-tuned policy interventions can hardly be expected to produce desirable outcomes in setting such as the latter. And to the extent that Washington Consensus policies are more conducive to honest behavior on the part of politicians, they may well be preferable on this account. However, the evidence is ambiguous on this. Most

[28] One month after taking power in a military coup in 1961, President Park arrested some of the leading businessmen in Korea under the newly passed Law for Dealing with Illicit Wealth Accumulation. These businessmen were subsequently set free under the condition that they establish new industrial firms and give up the shares to the government (Amsden 1989, 72).

policies, including those of the Washington Consensus type, are corruptible if the underlying political economy permits or encourages it. Consider for example Russia's experiment with mass privatization. It is widely accepted that this process was distorted and de-legitimized by asset grabs on the part of politically well-connected insiders. Washington Consensus policies themselves cannot legislate powerful rent-seekers out of existence. Rank ordering different policy regimes requires a more fully specified model of political economy than the reduced-form view that automatically associates governmental restraint with less rent-seeking.[29]

I close this section with the usual refrain: the range of strategies that have worked in the past is quite diverse. Traditional import-substituting industrialization (ISI) model was quite effective in stimulating growth in a large number of developing countries (e.g., Brazil, Mexico, Turkey). So was East Asian style outward orientation, which combined heavy-handed interventionism at home with single-minded focus on exports (South Korea, Taiwan). Chile's post-1983 strategy was based on quite a different style of outward orientation, relying on large real depreciation, absence of explicit industrial policies (but quite a bit of support for non-traditional exports in agro-industry), saving mobilization through pension privatization, and discouragement of short-term capital inflows. The experience of countries such as China and Mauritius is best described as two-track reform. India comes as close to genuine gradualism as one can imagine. Hong Kong represents probably the only case where growth has taken place without an active policy of crowding in private investment and entrepreneurship, but here too special and favorable preconditions (mentioned earlier) limit its relevance to other settings. In view of this diversity, any statement on what ignites growth has to be cast at a sufficiently high level of generality.

2. An institution building strategy to sustain growth

In the long run, the main thing that ensures convergence with the living standards of advanced countries is the acquisition of high-quality institutions. The growth-spurring strategies described above have to be complemented over time with a cumulative process of institution building to ensure that growth does not run out of steam and that the economy remains resilient to shocks. This point has now been amply demonstrated both by historical accounts (North and Thomas 1973, Engerman and Sokoloff 1994) and by econometric studies (Hall and Jones 1999, Acemoglu et al. 2001, Rodrik et al., 2002, Easterly and Levine, 2002). However, these studies tend to remain at a very aggregate level of generality and do not provide much policy guidance (a point that is also made in Besley and Burgess 2002b).

[29] In Rodrik (1995) I compared export subsidy regimes in six countries, and found that the regimes that were least likely to be open to rent-seeking ex ante—those with clear-cut rules, uniform schedules, and no arms' length relationships between firms and bureaucrats—were in fact less effective ex post. Where bureaucrats were professional and well-monitored, discretion was not harmful. Where they were not, the rules did not help.

The empirical research on national institutions has generally focused on the protection of property rights and the rule of law. But one should think of institutions along a much wider spectrum. In its broadest definition, institutions are the prevailing rules of the game in society (North 1990). High quality institutions are those that induce socially desirable behavior on the part of economic agents. Such institutions can be both informal (e.g., moral codes, self-enforcing agreements) and formal (legal rules enforced through third parties). It is widely recognized that the relative importance of formal institutions increases as the scope of market exchange broadens and deepens. One reason is that setting up formal institutions requires high fixed costs but low marginal costs, whereas informal institutions have high marginal costs (Li 1999; Dixit 2004, chap. 3). I will focus here on formal institutions.

What kind of institutions matter and why? Table 11 provides a taxonomy of market-sustaining institutions, associating each type of institutions with a particular need. The starting point is the recognition that markets need not be self-creating, self-regulating, self-stabilizing, and self-legitimizing. Hence, the very existence of market exchange presupposes property rights and some form of contract enforcement. This is the aspect of institutions that has received the most scrutiny in empirical work. The central dilemma here is that a political entity that is strong enough to establish property rights and enforce contracts is also strong enough, by definition, to violate these same rules for its own purpose (Djankov et al., 2003). The relevant institutions must strike the right balance between disorder and dictatorship.

As Table 11 makes clear, there are other needs as well. Every advanced economy has discovered that markets require extensive regulation to minimize abuse of market power, internalize externalities, deal with information asymmetries, establish product and safety standards, and so on. They also need monetary, fiscal, and other arrangements to deal with the business cycle and the problems of unemployment/inflation that are at the center of macroeconomists' analyses since Keynes. Finally, market outcomes need to be legitimized through social protection, social insurance, and democratic governance most broadly (Rodrik 2000).

Institutional choices made in dealing with these challenges often have to strike a balance between competing objectives. The regulatory regime governing the employment relationship must trade off the gains from "flexibility" against the benefits of stability and predictability. The corporate governance regime must delineate the interests and prerogatives of shareholders and stakeholders. The financial system must be free to take risks, but not so much so that it becomes an implicit public liability. There must be enough competition to ensure static allocative efficiency, but also adequate prospect of rents to spur innovation.

The last two centuries of economic history in today's rich countries can be interpreted as an ongoing process of learning how render capitalism more productive by supplying the institutional ingredients of a self-sustaining market economy: meritocratic public bureaucracies, independent judiciaries, central banking, stabilizing fiscal policy, antitrust and regulation, financial supervision, social insurance, political democracy. Just

as it is silly to think of these as the prerequisites of economic growth in poor countries, it is equally silly not to recognize that such institutions eventually become necessary to achieve full economic convergence. In this connection, one may want to place special emphasis on democratic institutions and civil liberties, not only because they are important in and of themselves, but also because they can be viewed as meta-institutions that help society make appropriate selections from the available menu of economic institutions.

However, the earlier warning not to confuse institutional function and institutional form becomes once again relevant here. Appropriate regulation, social insurance, macroeoconomic stability and the like can be provided through diverse institutional arrangements. While one can be sure that some types of arrangements are far worse than others, it is also the case that many well-performing arrangements are functional equivalents. Function does not map uniquely into form. It would be hard to explain otherwise how social systems that are so different in their institutional details as those of the United States, Japan, and Europe have managed to generate roughly similar levels of wealth for their citizens. All these societies protect property rights, regulate product, labor, and financial markets, have sound money, and provide for social insurance. But the rules of the game that prevail in the American style of capitalism are very different from those in the Japanese style of capitalism. Both differ from the European style. And even within Europe, there are large differences between the institutional arrangements in, say, Sweden and Germany. There has been only modest convergence among these arrangements in recent years, with the greatest amount of convergence taking place probably in financial market practices and the least in labor market institutions (Freeman 2000).

There are a number of reasons for institutional non-convergence. First, differences in social preferences, say over the tradeoff between equity and opportunity, may result in different institutional choices. If Europeans have a much greater preference for stability and equity than Americans, their labor market and welfare-state arrangements will reflect that preference. Second, complementarities among different parts of the institutional landscape can generate hysteresis and path dependence. An example of this would be the complementarity between corporate governance and financial market practices of the Japanese "model," as discussed previously. Third, the institutional arrangements that are required to promote economic development can differ significantly, both between rich and poor countries and among poor countries. This too has been discussed previously.

There is increasing recognition in the economics literature that high-quality institutions can take a multitude of forms and that economic convergence need not necessarily entail convergence in institutional forms (North 1994, Freeman 2000, Pistor 2000, Mukand and Rodrik forthcoming, Berkowitz et al. 2003, Djankov et al. 2003, Dixit 2004).[30] North (1994, 8) writes: "economies that adopt the formal rules of another

[30] Furthermore, as Roberto Unger (1998) has argued, there is no reason to suppose that today's advanced economies have already exhausted all the useful institutional variations that could underpin healthy and vibrant economies.

economy will have very different performance characteristics than the first economy because of different informal norms and enforcement [with the implication that] transferring the formal political and economic rules of successful Western economies to third-world and Eastern European economies is not a sufficient condition for good economic performance." Freeman (2000) discusses the variety of labor market institutions that prevail among the advanced countries and argues that differences in these practices have first-order distributional effects, but only second-order efficiency effects. Pistor (2000) provides a general treatment of the issue of legal transplantation, and shows how importation of laws can backfire. In related work, Berkowitz et al. (2003) find that countries that developed their formal legal orders internally, adapted imported codes to local conditions, or had familiarity with foreign codes ended up with much better legal institutions than those that simply transplanted formal legal orders from abroad. Djankov et al. (2003) base their discussion on an "institutional possibility frontier" that describes the tradeoff between private disorder and dictatorship, and argue that different circumstances may call for different choices along this frontier. And Dixit (2004, 4) summarizes the lessons for developing countries thus: "it is not always necessary to create replicas of western style state legal institutions from scratch; it may be possible to work with such alternative institutions as are available, and build on them."

Mukand and Rodrik (forthcoming) develop a formal model to examine the costs and benefits of institutional "experimentation" versus "copycatting" when formulas that have proved successful elsewhere may be unsuitable at home. A key idea is that institutional arrangements that prove successful in one country create both positive and negative spillovers for other countries. On the positive side, countries whose underlying conditions are sufficiently similar to those of the successful "leaders" can imitate the arrangements prevailing there and forego the costs of experimentation. This is one interpretation of the relative success that transition economies in the immediate vicinity of the European Union have experienced. Countries such as Poland, the Czech Republic or the Baltic republics share a similar historical trajectory with the rest of Europe, have previous experience with capitalist market institutions, and envisaged full EU membership within a reasonable period (de Menil 2003). The wholesale adoption of EU's acquis communautaire may have been the appropriate institution-building strategy for these countries. On the other hand, countries may be tempted or forced to imitate institutional arrangements for political or other reasons, even when their underlying conditions are too dissimilar for the strategy to make sense.[31] Institutional copycatting may have been useful for Poland, but it is much less clear that it was relevant or practical for Ukraine or Kyrgyzstan. The negative gradient in the economic performance of transition economies as one moves away from Western Europe provides some support for this idea (see Mukand and Rodrik forthcoming).

---

[31] In Mukand and Rodrik (forthcoming) it is domestic politics that generates inefficient imitation. Political leaders may want to signal their type (and increase the probability of remaining in power) by imitating standard policies even when they know these will not work as well as alternative arrangements. But one can also appeal to the role of IMF and World Bank conditionality in producing this kind of outcome.

Even though it is recent, this literature opens up a new and exciting way of looking at institutional reform. In particular, it promises an approach that is less focused on so-called best practices or the superiority of any particular model of capitalism, and more cognizant of the context-specificity of desirable institutional arrangements. Dixit's (2004) monograph outlines a range of theoretical models that help structure our thinking along these lines.

VI. Concluding remarks

Richard Feynman, the irreverent physicist who won the Nobel Prize in 1965 for his work on quantum electrodynamics, relates the following story. Following the award ceremony and the dinner in Stockholm, he wanders into a room where a Scandinavian princess is holding court. The princess recognizes him as one of the awardees and asks him what he got the prize for. When Feynman replies that his field is physics, the princess says that this is too bad. Since no one at the table knows anything about physics, she says, they cannot talk about it. Feynman disagrees:

> "On the contrary," I answered. "It's because somebody knows *something* about it that we can't talk about physics. It's the things that nobody knows anything about that we *can* discuss. We can talk about the weather; we can talk about social problems; we can talk about psychology; we can talk about international finance … so it's the subject that nobody knows anything about that we can all talk about!" (Feynman 1985)

This is not the place to defend international finance (circa 1965) against the charge Feynman levels at it. But suppose Feynman had picked on economic growth instead of international finance. Would growth economists have a plausible riposte? Is the reason we all talk so much about growth that we understand so little about it?

It is certainly the case that growth theory is now a much more powerful tool than it was before Solow put pencil to paper. And cross-country regressions have surely thrown out some useful correlations and stylized facts. But at least at the more practical end of things—how do we make growth happen? —things have turned out to be somewhat disappointing. By the mid-1980s, policy oriented economists had converged on a new consensus regarding the policy framework for growth. We thought we knew a lot about what governments needed to do. But as my Martian thought experiment at the beginning of the paper underscores, reality has been unkind to our expectations. If Latin America was booming today and China and India were stagnating, we would have an easier time fitting the world to our policy framework. Instead, we are straining to explain why unorthodox, two-track, gradualist reform paths have done so much better than sure-fire adoption of the standard package.

Very few policy analysts think that the answer is to go back to old-style ISI, even though its record was certainly respectable for a very large number of countries. Certainly no-one one believes that central planning is a credible alternative. But by the

same token, few are now convinced that liberalization, deregulation, and privatization on their own hold the key to unleashing economic growth. Maybe the right approach is to give up looking for "big ideas" altogether (as argued explicitly by Lindauer and Pritchett 2002, and implicitly by Easterly 2001). But that would be overshooting too. Economics is full of big ideas on the importance of incentives, markets, budget constraints, and property rights. It offers powerful ways of analyzing the allocative and distributional consequences of proposed policy changes. The key is to realize that these principles do not translate directly into specific policy recommendations. That translation requires the analyst to supply many additional ingredients that are contingent on the economic and political context, and cannot be done a priori. Local conditions matter not because economic principles change from place to place, but because those principles come institution free and filling them out requires local knowledge.

Therefore, the real lesson for the architects of growth strategies is to take economics more seriously, not less seriously. But the relevant economics is that of the seminar room, with its refusal to make unconditional generalizations and its careful examination of the contingent relation between the economic environment and policy implications. Rule-of-thumb economics, which has long dominated thinking on growth policies, can be safely discarded.

Acknowledgements

I gratefully acknowledge financial support from the Carnegie Corporation of New York. I also thank, without implicating, Philippe Aghion, Richard Freeman, Steph Haggard, Ricardo Hausmann, Murat Iyigun, Sharun Mukand, José Antonio Ocampo, Andrei Shleifer, and Arvind Subramanian for comments that substantially improved this paper.

REFERENCES

Acemoglu, Daron, Simon Johnson, and James A. Robinson, (2001), "The Colonial Origins of Comparative Development: An Empirical Investigation," American Economic Review 91(5): December 1369-1401.

Acemoglu, Daron, Philippe Aghion, and Fabrizio Zilibotti, (2002), "Distance to Frontier, Selection, and Economic Growth," NBER Working Paper No. 9066, July.

Aghion, Philippe, Robin Burgess, Stephen Redding and Fabrizio Zilibotti, (2003)"The Unequal Effects of Liberalization: Theory and Evidence from India," Department of Economics, London School of Economics, March.

Amsden, Alice H., (1989),Asia's Next Giant: South Korea and Late Industrialization, Oxford University Press (New York and Oxford).

Anderson, James E., and J. Peter Neary, (1992) "Trade Reform with Quotas, Partial Rent Retention, and Tariffs," Econometrica 60: 57-76.

Aoki, Masahiko, (1997), "Unintended Fit: Organizational Evolution and Government Design of Institutions in Japan," in M. Aoki et al, eds., The Role of Government in East Asian Economic Development: Comparative Institutional Analysis, (Clarendon Press, Oxford).

Aslund, Anders, Peter Boone, and Simon Johnson, (1996) "How to Stabilize: Lessons from Post-Communist Countries," Brookings Papers on Economic Activity 1.

Aslund, Anders, and Simon Johnson, (2003) "Small Enterprises and Economic Policy," Working Paper, Sloan School, MIT.

Azariadis C., and A. Drazen (1990) "Threshold Externalities in Economic Development" Quarterly Journal of Economics 105, 501–526.

Berkowitz, Daniel, Katharina Pistor, and Jean-Francois Richard, (2003) "Economic Development, Legality, and the Transplant Effect," European Economic Review, 47(1):165-195.

Besley, Timothy, and Robin Burgess, (2002) "Can Labor Regulation Hinder Economic Performance? Evidence from India," CEPR Discussion Paper No. 3260.

Besley, Timothy, and Robin Burgess, (2002) "Halving Global Poverty," Department of Economics, London School of Economics, August. (2002b)

Birdsall, Nancy, and Augusto de la Torre, (2001) "Washington Contentious: Economic Policies for Social Equity in Latin America." Washington: Carnegie Endowment for International Peace and Inter-American Dialogue.

Bosworth, Barry, and Susan M. Collins, (2003)"The Empirics of Growth: An Update," Brookings Institutions, unpublished paper, March 7, 2003.

Brock, William A., and Steven N. Durlauf, (2001) "Growth Empirics and Reality," The World Bank Economic Review, 15(2): 229-272.

Calvo, Guillermo, "Incredible Reforms," (1989) in Calvo *et al*. eds., Debt, Stabilization and Development, (New York, Basil Blackwell).

Collins, Susan, and Barry Bosworth, (1996) "Economic Growth in East Asia: Accumulation versus Assimilation," Brookings Papers on Economic Activity 1996:2, 135-191.

DeLong, Brad, (2003) "India since Independence: An Analytic Growth Narrative," in Dani Rodrik, ed., In Search of Prosperity: Analytic Narratives of Economic Growth, (Princeton University Press, Princeton, NJ).

De Menil, Georges, (2003) "History, Policy, and Performance in Two Transition Economies: Poland and Romania," in Dani Rodrik, ed., In Search of Prosperity: Analytic Narratives of Economic Growth, (Princeton University Press, Princeton, NJ).

Dewatripont, Mathias, and Gerard Roland, (1995) "The Design of Reform Packages under Uncertainty," American Economic Review, 85(5): 1207-23.

Dixit, Avinash, (2004) Lawlessness and Economics: Alternative Modes of Economic Governance, Gorman Lectures (forthcoming, Princeton University Press).

Djankov, Simeon, Edward Glaeser, Rafael LaPorta, Florencio Lopez-de-Silanes, and Andrei Shleifer, "The New Comparative Economics," Harvard University, January 2003.

Easterly, William, (2001) The Elusive Quest for Growth, (MIT Press, Cambridge, MA).

Easterly, William, (2003) "National Policies and Economic Growth: A Reappraisal," New York University, Development Research Institute (DRI) Working Paper No. 1, March 2003.

Easterly, William, Michael Kremer, Lant Pritchett and Lawrence H. Summers, (1993) "Good Policy or Good Luck? Country Growth Performance and Temporary Shocks," Journal of Monetary Economics 32(3): 459-483.

Easterly, W., and R. Levine, (2002) "Tropics, Germs, and Crops: How Endowments Influence Economic Development," mimeo, Center for Global Development and Institute for International Economics.

Engerman, Stanley L., and Kenneth L. Sokoloff, (1994) "Factor Endowments, Institutions, and Differential Paths of Growth Among New World Economies: A View from Economic Historians of the United States," National Bureau of Economic Research Working Paper No. H0066, December.

Feynman, Richard P., (1985) "Surely You're Joking Mr. Feynman!", W.W. Norton, New York)

Frankel, Jeffrey, (2000) "The Asian Model, the Miracle, the Crisis, and the Fund," in P. Krugman, ed., Currency Crises, The University of Chicago Press for the NBER.

Freeman, Richard B., (2000) "Single Peaked vs. Diversified Capitalism: The Relation Between Economic Institutions and Outcomes," NBER Working Paper No. W7556, February.

Friedman, Eric, Simon Johnson, Daniel Kaufmann, and Pablo Zoido-Lobaton, (2000) "Dodging the Grabbing Hand: The Determinants of Unofficial Activity in 69 Countries," Journal of Public Economics, 76: 459-493.

Gerschenkron, Alexander, (1962) Economic Backwardness in Historical Perspective: A Book of Essays, (Harvard University Press, Cambridge, MA).

Gilbert, Christopher L., and Panos Varangis, (2003) "Globalization and International Commodity Trade with Specific Reference to West African Cocoa Producers," NBER Working Paper No. w9668, May.

Goldstone, Jack A., The Happy Chance: The Rise of the West in Global Context, 1500-1850, book manuscript in preparation, U.C. Davis, forthcoming.

Haggard, Stephan, (2003) "Institutions and Growth in East Asia," UCSD, unpublished manuscript.

Haggard, Stephan, and Robert Kaufman, eds., (1983) The Politics of Economic Adjustment, (Princeton University Press, Princeton, NJ).

Hall, Robert, and Chad I. Jones, (1999) "Why Do Some Countries Produce So Much More Output per Worker than Others?" Quarterly Journal of Economics, 114(1): 83-116.

Harberger, Arnold C., (1985) Economic Policy and Economic Growth, International Center for Economic Growth, Institute for Contemporary Studies, San Francisco, CA.

Harberger, Arnold C., "Interview with Arnold Harberger: Sound Policies Can Free Up Natural Forces of Growth," IMF Survey, International Monetary Fund, Washington, DC, July 14, 2003, 213-216.

Harrison, Glenn W., Thomas F. Rutherford, and David G . Tarr, (1993) "Trade Reform in the Partially Liberalized Economy of Turkey," The World Bank Economic Review, 7 (2), 191-218.

Hatta, Tatsuo, (1977) "A Recommendation for a Better Tariff Structure," Econometrica 45: 1859-69.

Hausmann, Ricardo and Dani Rodrik, (2002) "Economic Development as Self-Discovery," NBER Discussion Paper No. w8952, May.

Hausmann, Ricardo, Lant Pritchett, and Dani Rodrik. (2004). "Growth Accelerations," NBER Working Paper No. w10566, June.

Helleiner, Gerald K., ed., (1994) Trade Policy and Industrialization in Turbulent Times, UNU/WIDER, (Routledge, New York).

Hellmann, Thomas, Kevin Murdock, and Joseph Stiglitz, (1997) "Financial Restraint: Toward a New Paradigm," in M. Aoki et al, eds., The Role of Government in East Asian Economic Development: Comparative Institutional Analysis, (Clarendon Press, Oxford).

Hirschman, Albert O., (1958) The Strategy of Economic Development, (Yale University Press, New Haven, CT).

Hoff, Karla and Joseph Stiglitz, (2001) "Modern Economic Theory and Development," in G.M. Meier and J.E. Stiglitz, eds., Frontiers of Development Economics, New York, Oxford University Press, 389-459.

Imbs, Jean, and Romain Wacziarg, (2003) "Stages of Diversification," American Economic Review, 93(1): 63-86.

Johnson, Simon, John McMillan, and Chris Woodruff, (2000) "Entrepreneurs and the Ordering of Institutional Reform: Poland, Slovakia, Romania, Russia, and Ukraine Compared," Economics of Transition.

Kaufmann, Daniel, (2002) "Rethinking Governance," World Bank Institute, World Bank, Washington, D.C., December.

Krueger, Anne O. , (1997) "Trade Policy and Development: How We Learn," The American Economic Review, March.

Krugman, Paul, (1995) "Dutch Tulips and Emerging Markets", Foreign Affairs, July/August.

Kuczynski, Pedro-Pablo, and John Williamson, eds., (2003) After the Washington Consensus: Restarting Growth and Reform in Latin America, Institute for International Economics, Washington, DC.

Lau, Lawrence, J., Yingyi Qian, and Gerard Roland, (2000) "Reform Without Losers: An Interpretation of China's Dual-Track Approach to Transition," The Journal of Political Economy, 108(1): 120-143.

Li, Shuhe, (1999) "The Benefits and Costs of Relation-Based Governance: An Explanation of the East Asian Miracle and Crisis," City University of Hong Kong, October.

Lin, Justin Yifu, Fang Cai, and Zhou Li, (1996) The China Miracle: Development Strategy and Economic Reform, The Chinese University Press, Shatin, NT, Hong Kong.

Lin, Justin Yifu, and Mingxing Liu, (2003) "Development Strategy, Viability and Challenges of Development in Lagging Regions," paper prepared for the 15th World Bank's Annual Bank Conference on Development Economics, Bangalore, India, May.

Lindauer, David L., and Lant Pritchett, (2002) "What's the Big Idea? The Third Generation of Policies for Economic Growth," Economia, 1-40.

Lipton, David, and Jeffrey Sachs, (1990) "Creating a Market Economy in Eastern Europe: The Case of Poland," Brookings Papers on Economic Activity, 1.

Loayza, Norman, Pablo Fajnzylber, and Cesar Calderon, (2002)"Economic Growth in Latin America and the Caribbean" Stylized Facts, Explanations, and Forecasts," World Bank, Washington, D.C., June.

Lopez, Ramon, (1997) "Environmental Externalities in Traditional Agriculture and the Impact of Trade Liberalization: The Case of Ghana," Journal of Development Economics, 53(1): 17-39.

Lopez, Ramon, and Arvind Panagariya, (1992)"On the Theory of Piecemeal Tariff Reform: The Case of Pure Imported Intermediate Inputs," American Economic Review, 82(3): 615-625.

Lora, Eduardo, (2001a) "Structural Reforms in Latin America: What Has Been Reformed and How to Measure It," Inter-American Development Bank, Washington, D.C., December.

Lora, Eduardo, (2001b)"El crecimiento económico en América Latina después de una década de reformas estructurales". Washington, DC, United States: Inter-American Development Bank, Research Department. Mimeographed document.

MacLean, Brian K., (1999)"The Rise and Fall of the 'Crony Capitalism' Hypothesis: Causes and Consequences," Department of Economics, Laurentian University, Ontario, March.

Maddison, Angus, (2001) The World Economy: A Millennial Perspective, OECD Development Centre, (OECD, Paris).

Matsuyama, Kiminori, (1992) "Agricultural Productivity, Comparative Advantage, and Economic Growth," Journal of Economic Theory, December: 317-334.

Milanovic, Branko, (2003) "The Two Faces of Globalization: Against Globalization as we Know It," World Development, 31(4): 667-683.

Mukand, Sharun, and Dani Rodrik, (forthcoming) "In Search of the Holy Grail: Policy Convergence, Experimentation, and Economic Performance," NBER Working Paper, January 2002 (American Economic Review, forthcoming).

Murphy, Kevin M., Andrei Shleifer, and Robert W. Vishny, (1989) "Industrialization and the Big Push," Journal of Political Economy, Vol. 97 (5): 1003-26.

Murphy, Kevin M., Andrei Shleifer, and Robert W. Vishny, (1992) "The Transition to a Market Economy: Pitfalls of Partial Reform," The Quarterly Journal of Economics, 107(3): 889-906.

Naim, Moises, (1999) "Fads and Fashion in Economic Reforms: Washington Consensus or Washington Confusion?" paper prepared for the IMF Conference on Second Generation Reforms, Washington, D.C., October.

North, Douglass C., (1990) Institutions, Institutional Change and Economic Performance, (Cambridge University Press, New York).

North, Douglass C., (1994)"Economic Performance Through Time," The American Economic Review, 84(3): 359-368.

North, Douglass C., and R. Thomas, (1973) The Rise of the Western World: A New Economic History, (Cambridge University Press, Cambridge).

Ocampo, José Antonio, (2002) "Rethinking the Development Agenda," United Nations Economic Commission for Latin America and the Caribbean (ECLAC), Santiago, Chile.

Ocampo, José Antionio, (2003) "Structural Dynamics and Economic Growth in Developing Countries," United Nations Economic Commission for Latin America and the Caribbean (ECLAC), Santiago, Chile.

Pistor, Katharina, (2000) "The Standardization of Law and its Effect on Developing Economies," G-24 Discussion Paper No. 4, July.

Polterovich, Victor, and Vladimir Popov, (2002), "Accumulation of Foreign Exchange Reserves and Long Term Growth," New Economic School, Moscow, Russia, unpublished paper.

Qian, Yingyi, (2003) “How Reform Worked in China,” in D. Rodrik, ed., In Search of Prosperity: Analytic Narratives of Economic Growth, Princeton, NJ, Princeton University Press.

Rodríguez, Francisco, and Dani Rodrik, (2001) "Trade Policy and Economic Growth: A Skeptic's Guide to the Cross-National Evidence," Macroeconomics Annual 2000, eds. Ben Bernanke and Kenneth S. Rogoff, (MIT Press for NBER, Cambridge, MA).

Rodriguez-Clare, Andres, "The Division of Labor and Economic Development," Journal of Development Economics, 49 (April), 3-32.

Rodrik, Dani, (1991) "Policy Uncertainty and Private Investment in Developing Countries," Journal of Development Economics 36, (November).

Rodrik, Dani, (1995) "Taking Trade Policy Seriously: Export Subsidization as a Case Study in Policy Effectiveness," in A. Deardorff, J. Levinson, and R. Stern (eds.), New Directions in Trade Theory, (University of Michigan Press, Ann Arbor).

Rodrik, Dani, (1996) "Coordination Failures and Government Policy: A Model with Applications to East Asia and Eastern Europe," Journal of International Economics 40(1-2): (February) 1-22.

Rodrik, Dani (1996)“Understanding Economic Policy Reform,” Journal of Economic Literature, XXXIV: (March) 9-41.

Rodrik, Dani, (1997) "Trade Strategy, Exports, and Investment: Another Look at East Asia," Pacific Economic Review, (February).

Rodrik, Dani, (1999) The New Global Economy and Developing Countries: Making Openness Work, Washington, D.C., Overseas Development Council.

Rodrik, Dani, (1999b) "Where Did All the Growth Go? External Shocks, Social Conflict and Growth Collapses," Journal of Economic Growth, (December).

Rodrik, Dani, (2000) "Institutions for High-Quality Growth: What They Are and How to Acquire Them," Studies in Comparative International Development, 35(3): Fall.

Rodrik, Dani, and Arvind Subramanian, (2004). “From ‘Hindu Growth’ to Productivity Surge: The Mystery of the Indian Growth Transition,” NBER Working Paper No. w10376, March.

Rodrik, Dani, Arvind Subramanian, and Francesco Trebbi, (2002) “Institutions Rule: The Primacy of Institutions over Geography and Integration in Economic Development” Kennedy School of Government, Harvard University, (October).

Rosenstein-Rodan, Paul, (1943) "Problems of Industrialization of Eastern and Southeastern Europe." Economic Journal 53(210-211): 202-211.

Rostow, Walt W., (1965) The Stages of Economic Growth: A Non-Communist Manifesto, (Cambridge University Press, Cambridge and New York).

Shleifer, Andrei, and Robert W. Vishny, (1998) The Grabbing Hand: Government Pathologies and Their Cures, (Harvard University Press, Cambridge, MA).

Soon, Cho, (1994) The Dynamics of Korean Development, Washington, DC, Institute for International Economics.

Stern, Nicholas, (2001)"A Strategy for Development," ABCDE Keynote Address, Washington, DC, World Bank, (May).

Stiglitz, Joseph E., (1998) "More Instruments and Broader Goals Moving toward the Post-Washington Consensus." United Nations University/WIDER, Helsinki.

Subramanian, Arvind, and Devesh Roy, (2003) "Who Can Explain the Mauritian Miracle? Meade, Romer, Sachs, or Rodrik?" in Dani Rodrik, ed., In Search of Prosperity: Analytic Narratives of Economic Growth, (Princeton University Press, Princeton, NJ).

Summers, Lawrence H., (2003) "Godkin Lectures," John F. Kennedy School of Government, Harvard University, (April).

Temple, Jonathan, (1999) "The New Growth Evidence," Journal of Economic Literature, 37(1): 112-156.

Temple, Jonathan, (2003) "Growing into Trouble: Indonesia since 1966," in Dani Rodrik, ed., In Search of Prosperity: Analytic Narratives of Economic Growth, (Princeton University Press, Princeton, NJ).

Trindade, Vitor, (2003) "The Big Push, Industrialization, and International Trade: The Role of Exports," Maxwell School, Syracuse University, (March).

Unger, Roberto Mangabeira, (1998) Democracy Realized: The Progressive Alternative, Verso, London and New York.

Vamvakidis, Athanasios, (2002) "How Robust is the Growth-Openness Connection? Historical Evidence," Journal of Economic Growth, 7(1): (March) 57-80.

Van Arkadie, Brian, and Raymond Mallon, (2003) Vietnam: A Transition Tiger?, Asia Pacific Press at The Australian National University, Australia.

Wei, Shang-Jin, (1997) "Gradualism versus Big Bang: Speed and Sustainability of Reforms," Canadian Journal of Economics, 30(4B): 1234-47.

Wellisz, Stanislaw, and Philippe Lam Shin Saw, (1993) "Mauritius," in Ronald Findlay and Stanislaw Wellisz, eds., The Political Economy of Poverty, Equity, and Growth: Five Open Economies, (New York, Oxford University Press).

Williamson, John, (1990)"What Washington Means by Policy Reform", in J. Williamson, ed., Latin American Adjustment: How Much Has Happened? (Washington: Institute for International Economics).

Williamson, John, and Roberto Zagha, (2002)"From Slow Growth to Slow Reform," World Bank, unpublished paper.

World Bank, (1993) The East Asian Miracle: Economic Growth and Public Policies, (World Bank, Washington, D.C.).

World Bank, (1998) Beyond the Washington Consensus: Institutions Matter, (World Bank, Washington, D.C.).

Yanikkaya, Halit, (2003) "Trade Openness and Economic Growth: A Cross-Country Empirical Investigation," Journal of Development Economics, 72(October): 57-89.

Young, Alwyn, (1992) "A Tale of Two Cities: Factor Accumulation and Technical Change in Hong Kong and Singapore," NBER Macroeconomics Annual, (MIT Press for NBER, Cambridge, MA).

Young, Alwyn, (2000) "The Razor's Edge: Distortions and Incremental Reform in the People's Republic of China," NBER Working Paper No. 7828, (August).

Table 1. Sources of growth by regions, 1960-2000 (percent increase)

			Contribution of:		
Region/Period	Output	Output per worker	Physical capital	Education	Productivity
World (84)					
1960-70	5.1	3.5	1.2	0.3	1.9
1970-80	3.9	1.9	1.1	0.5	0.3
1980-90	3.5	1.8	0.8	0.3	0.8
1990-2000	3.3	1.9	0.9	0.3	0.8
Industrial Countries (22)					
1960-70	5.2	3.9	1.3	0.3	2.2
1970-80	3.3	1.7	0.9	0.5	0.3
1980-90	2.9	1.8	0.7	0.2	0.9
1990-2000	2.5	1.5	0.8	0.2	0.5
China (1)					
1960-70	2.8	0.9	0.0	0.3	0.5
1970-80	5.3	2.8	1.6	0.4	0.7
1980-90	9.2	6.8	2.1	0.4	4.2
1990-2000	10.1	8.8	3.2	0.3	5.1
East Asia less China (7)					
1960-70	6.4	3.7	1.7	0.4	1.5
1970-80	7.6	4.3	2.7	0.6	0.9
1980-90	7.2	4.4	2.4	0.6	1.3
1990-2000	5.7	3.4	2.3	0.5	0.5
Latin America (22)					
1960-70	5.5	2.8	0.8	0.3	1.6
1970-80	6.0	2.7	1.2	0.3	1.1
1980-90	1.1	-1.8	0.0	0.5	-2.3
1990-2000	3.3	0.9	0.2	0.3	0.4
South Asia (4)					
1960-70	4.2	2.2	1.2	0.3	0.7
1970-80	3.0	0.7	0.6	0.3	-0.2
1980-90	5.8	3.7	1.0	0.4	2.2
1990-2000	5.3	2.8	1.2	0.4	1.2
Africa (19)					
1960-70	5.2	2.8	0.7	0.2	1.9
1970-80	3.6	1.0	1.3	0.1	-0.3
1980-90	1.7	-1.1	-0.1	0.4	-1.4
1990-2000	2.3	-0.2	-0.1	0.4	-0.5
Middle East (9)					
1960-70	6.4	4.5	1.5	0.3	2.6
1970-80	4.4	1.9	2.1	0.5	-0.6
1980-90	4.0	1.1	0.6	0.5	0.1
1990-2000	3.6	0.8	0.3	0.5	0.0

Source: Bosworth and Collins (2003).

Table 2: Rules of good behavior for promoting economic growth

Original Washington Consensus:	"Augmented" Washington Consensus: … the previous 10 items, plus:
1. Fiscal discipline 2. Reorientation of public expenditures 3. Tax reform 4. Interest rate liberalization 5. Unified and competitive exchange rates 6. Trade liberalization 7. Openness to DFI 8. Privatization 9. Deregulation 10. Secure Property Rights	11. Corporate governance 12. Anti-corruption 13. Flexible labor markets 14. Adherence to WTO disciplines 15. Adherence to international financial codes and standards 16. "Prudent" capital-account opening 17. Non-intermediate exchange rate regimes 18. Independent central banks/inflation targeting 19. Social safety nets 20. Targeted poverty reduction

Table 3: East Asian anomalies

Institutional domain	**Standard ideal**	**"East Asian" pattern**
Property rights	Private, enforced by the rule of law	Private, but govt authority occasionally overrides the law (esp. in Korea).
Corporate governance	Shareholder ("outsider") control, protection of shareholder rights	Insider control
Business-government relations	Arms' length, rule based	Close interactions
Industrial organization	Decentralized, competitive markets, with tough anti-trust enforcement	Horizontal and vertical integration in production (chaebol); government-mandated "cartels"
Financial system	Deregulated, securities based, with free entry. Prudential supervision through regulatory oversight.	Bank based, restricted entry, heavily controlled by government, directed lending, weak formal regulation.
Labor markets	Decentralized, de-institutionalized, "flexible" labor markets	Lifetime employment in core enterprises (Japan)
International capital flows	"prudently" free	Restricted (until the 1990s)
Public ownership	None in productive sectors	Plenty in upstream industries.

Table 4: The logic of the Washington Consensus and a Chinese counterfactual

Problem		Solution
Low agricultural productivity	→	Price liberalization
Production incentives	→	Land privatization
Loss of fiscal revenues	→	Tax reform
Urban wages	→	Corporatization
Monopoly	→	Trade liberalization
Enterprise restructuring	→	Financial sector reform
Unemployment	→	Social safety nets
… and so on		

Table 5: How to understand/rationalize institutional anomalies: four illustrations

Objective	**What is problem?**	**Institutional response**	**Prerequisites**	**Institutional complements**
Financial deepening (saving mobilization and efficient intermediation)	Asymmetric information (investors know more about their projects than lenders do) and limited liability	"Financial restraint" (Hellmann et al. 1997): controlled deposit rates and restricted entry —creation of rents to induce better portfolio risk management, better monitoring of firms, and increased deposit mobilization by banks.	Ability to maintain restraint at <u>moderate</u> levels; Positive real interest rates; Macroeconomic stability; Avoid state capture by financial interests.	<u>Finance</u>: Highly regulated financial markets (absence of security markets and closed capital accounts to prevent cherry picking and rent dissipation); <u>Politics</u>: State "autonomy" to prevent capture and decay into "crony capitalism."

Table 5: How to understand/rationalize institutional anomalies: four illustrations (cont.)

Objective	**What is problem?**	**Institutional response**	**Prerequisites**	**Institutional complements**
Spurring investment and entrepreneurship in non-traditional activities	Economies of scale together with inter-industry linkages depress private return to entrepreneurship/investment below social return.	"Industrial policy as a coordination device" (Rodrik 1995) --credit subsidies (Korea) and tax incentives (Taiwan) for selected sectors; --protection of home market coupled with export subsidies; --public enterprise creation for upstream products; --arm-twisting and cajoling by political leadership; --socialization of investment risk through implicit investment guarantees.	A high level of human capital relative to physical capital. A relatively competent bureaucracy to select investment projects.	Trade: Need to combine import protection (in selected sectors) with exposure to competition in export markets to distinguish high-productivity firms from low-productivity ones; Business-government relations: "Embedded autonomy" (Evans) to enable close interactions and information exchange while preventing state capture and decay into "crony capitalism."

Table 5: How to understand/rationalize institutional anomalies: four illustrations (cont.)

Objective	**What is problem?**	**Institutional response**	**Prerequisites**	**Institutional complements**
Productive organization of the workplace	Tradeoff between information sharing (working together) and economies of specialization (specialized tasks)	"horizontal hierarchy" (Aoki 1997)	(unintended) fit with prewar arrangements of military resource mobilization in Japan	Corporate governance: insider control to provide incentive for accumulating long-term managerial skills; Labor markets: lifetime employment and enterprise unionism to generate long-term collaborative teamwork; Financial markets: main bank system to discipline firms and reduce the moral hazard consequences of insider control; Politics: "bureau-pluralism" (regulation, protection) to redistribute benefits to less productive, traditional sectors.

Table 5: How to understand/rationalize institutional anomalies: four illustrations (cont.)

Objective	**What is problem?**	**Institutional response**	**Prerequisites**	**Institutional complements**
Reduce anti-export bias	Import-competing interests are politically powerful and opposed to trade liberalization	export processing zone (Rodrik 1999)	saving boom; elastic supply of foreign investment; preferential market access in EU	Dual labor markets: segmentation between male and female labor force, so that increase female employment in the EPZ does not drive wages up in the rest of the economy.

Table 6: Sound economics and institutional counterparts: microeconomics

OBJECTIVE	**UNIVERSAL PRINCIPLES**	**PLAUSIBLE DIVERSITY IN INSTITUTIONAL ARRANGEMENTS**
Productive efficiency (static and dynamic)	Property rights: Ensure potential and current investors can retain the returns to their investments Incentives: Align producer incentives with social costs and benefits. Rule of law: Provide a transparent, stable and predictable set of rules.	What type of property rights? Private, public, cooperative? What type of legal regime? Common law? Civil law? Adopt or innovate? What is the right balance between decentralized market competition and public intervention? Which types of financial institutions/corporate governance are most appropriate for mobilizing domestic savings? Is there a public role to stimulate technology absorption and generation? (e.g. IPR "protection")

Table 7: Sound economics and institutional counterparts: macroeconomics

OBJECTIVE	**UNIVERSAL PRINCIPLES**	**PLAUSIBLE DIVERSITY IN INSTITUTIONAL ARRANGEMENTS**
Macroeconomic and Financial Stability	Sound money: Do not generate liquidity beyond the increase in nominal money demand at reasonable inflation. Fiscal sustainability: Ensure public debt remains "reasonable" and stable in relation to national aggregates. Prudential regulation: Prevent financial system from taking excessive risk.	How independent should the central bank be? What is the appropriate exchange-rate regime? (dollarization, currency board, adjustable peg, controlled float, pure float) Should fiscal policy be rule-bound, and if so what are the appropriate rules? Size of the public economy. What is the appropriate regulatory apparatus for the financial system? What is the appropriate regulatory treatment of capital account transactions?

Table 8: Sound economics and institutional counterparts: social policy

OBJECTIVE	**UNIVERSAL PRINCIPLES**	**PLAUSIBLE DIVERSITY IN INSTITUTIONAL ARRANGEMENTS**
Distributive justice and poverty alleviation	Targeting: Redistributive programs should be targeted as closely as possible to the intended beneficiaries. Incentive compatibility: Redistributive programs should minimize incentive distortions.	How progressive should the tax system be? Should pension systems be public or private? What are the appropriate points of intervention: educational system? access to health? access to credit? labor markets? tax system? What is the role of "social funds"? Redistribution of endowments? (land reform, endowments-at-birth) Organization of labor markets: decentralized or institutionalized? Modes of service delivery: NGOs, participatory arrangements., etc.

Table 9: Episodes of rapid growth, by region, decade and magnitude of acceleration						
Region	Decade	Country	Year	Growth before	Growth after	Difference in growth
Sub-Saharan Africa	1950s and 1960s	NGA	1967	-1.7	7.3	9.0
		BWA	1969	2.9	11.7	8.8
		GHA	1965	-0.1	8.3	8.4
		GNB	1969	-0.3	8.1	8.4
		ZWE	1964	0.6	7.2	6.5
		COG	1969	0.9	5.4	4.5
		NGA	1957	1.2	4.3	3.0
	1970s	MUS	1971	-1.8	6.7	8.5
		TCD	1973	-0.7	7.3	8.0
		CMR	1972	-0.6	5.3	5.9
		COG	1978	3.1	8.2	5.1
		UGA	1977	-0.6	4.0	4.6
		LSO	1971	0.7	5.3	4.6
		RWA	1975	0.7	4.0	3.3
		MLI	1972	0.8	3.8	3.0
		MWI	1970	1.5	3.9	2.5
	1980s and 1990s	GNB	1988	-0.7	5.2	5.9
		MUS	1983	1.0	5.5	4.4
		UGA	1989	-0.8	3.6	4.4
		MWI	1992	-0.8	4.8	5.6
South Asia	1950s/60s	PAK	1962	-2.4	4.8	7.1
	1970s	PAK	1979	1.4	4.6	3.2
		LKA	1979	1.9	4.1	2.2
	1980s	IND	1982	1.5	3.9	2.4
East Asia	1950s and 1960s	THA	1957	-2.5	5.3	7.8
		KOR	1962	0.6	6.9	6.3
		IDN	1967	-0.8	5.5	6.2
		SGP	1969	4.2	8.2	4.0
		TWN	1961	3.3	7.1	3.8
	1970s	CHN	1978	1.7	6.7	5.1
		MYS	1970	3.0	5.1	2.1
	1980s and 1990s	MYS	1988	1.1	5.7	4.6
		THA	1986	3.5	8.1	4.6
		PNG	1987	0.3	4.0	3.7
		KOR	1984	4.4	8.0	3.7
		IDN	1987	3.4	5.5	2.1
		CHN	1990	4.2	8.0	3.8

Table 9 (cont.): Episodes of rapid growth, by region, decade and magnitude of acceleration						
Region	Decade	Country	Year	Growth before	Growth after	Difference in growth
Latin America and Caribbean	1950s and 1960s	DOM	1969	-1.1	5.5	6.6
		BRA	1967	2.7	7.8	5.1
		PER	1959	0.8	5.2	4.4
		PAN	1959	1.5	5.4	3.9
		NIC	1960	0.9	4.8	3.8
		ARG	1963	0.9	3.6	2.7
		COL	1967	1.6	4.0	2.4
	1970s	ECU	1970	1.5	8.4	6.8
		PRY	1974	2.6	6.2	3.7
		TTO	1975	1.9	5.4	3.5
		PAN	1975	2.6	5.3	2.7
		URY	1974	1.5	4.0	2.6
	1980s and 1990s	CHL	1986	-1.2	5.5	6.7
		URY	1989	1.6	3.8	2.1
		HTI	1990	-2.3	12.7	15.0
		ARG	1990	-3.1	6.1	9.2
		DOM	1992	0.4	6.3	5.8
Middle East and North Africa	1950s and 1960s	MAR	1958	-1.1	7.7	8.8
		SYR	1969	0.3	5.8	5.5
		TUN	1968	2.1	6.6	4.5
		ISR	1967	2.8	7.2	4.4
		ISR	1957	2.2	5.3	3.1
	1970s	JOR	1973	-3.6	9.1	12.7
		EGY	1976	-1.6	4.7	6.3
		SYR	1974	2.6	4.8	2.2
		DZA	1975	2.1	4.2	2.1
	1980s and 1990s	SYR	1989	-2.9	4.4	7.3
OECD	1950s and 1960s	ESP	1959	4.4	8.0	3.5
		DNK	1957	1.8	5.3	3.5
		JPN	1958	5.8	9.0	3.2
		USA	1961	0.9	3.9	3.0
		CAN	1962	0.6	3.6	2.9
		IRL	1958	1.0	3.7	2.7
		BEL	1959	2.1	4.5	2.4
		NZL	1957	1.5	3.8	2.4
		AUS	1961	1.5	3.8	2.3
		FIN	1958	2.7	5.0	2.2
		FIN	1967	3.4	5.6	2.2
	1980s and 1990s	PRT	1985	1.1	5.4	4.3
		ESP	1984	0.1	3.8	3.7
		IRL	1985	1.6	5.0	3.4
		GBR	1982	1.1	3.5	2.5
		FIN	1992	1.0	3.7	2.8
		NOR	1991	1.4	3.7	2.2

Source: Hausmann et al. (2004).

Table 10: A taxonomy of "natural" barriers to industrialization

A. Learning externalities

    1. Learning-by-doing (e.g., Matsuyama, 1992)

    2. Human capital externalities (e.g., Azariadis and Drazen, 1990)

    3. Learning about costs (e.g., Hausmann and Rodrik, 2002)

B. Coordination failures (market-size externalities induced by IRS)

    1. Wage premium in manufacturing (e.g., Murphy, Shleifer, and Vishny, 1989)

    2. Infrastructure (e.g., Murphy, Shleifer, and Vishny, 1989)

    3. Specialized intermediate inputs (e.g., Rodrik 1994, 1995)

    4. Spillovers associated with wealth distribution (e.g., Hoff and Stiglitz 2001)

Table 11: A taxonomy of market-sustaining institutions

- Market-creating institutions
  - Property rights
  - Contract enforcement
- Market-regulating institutions
  - Regulatory bodies
  - Other mechanisms for correcting market failures
- Market-stabilizing institutions
  - Monetary and fiscal institutions
  - Institutions of prudential regulation and supervision
- Market-legitimizing institutions
  - Democracy
  - Social protection and social insurance

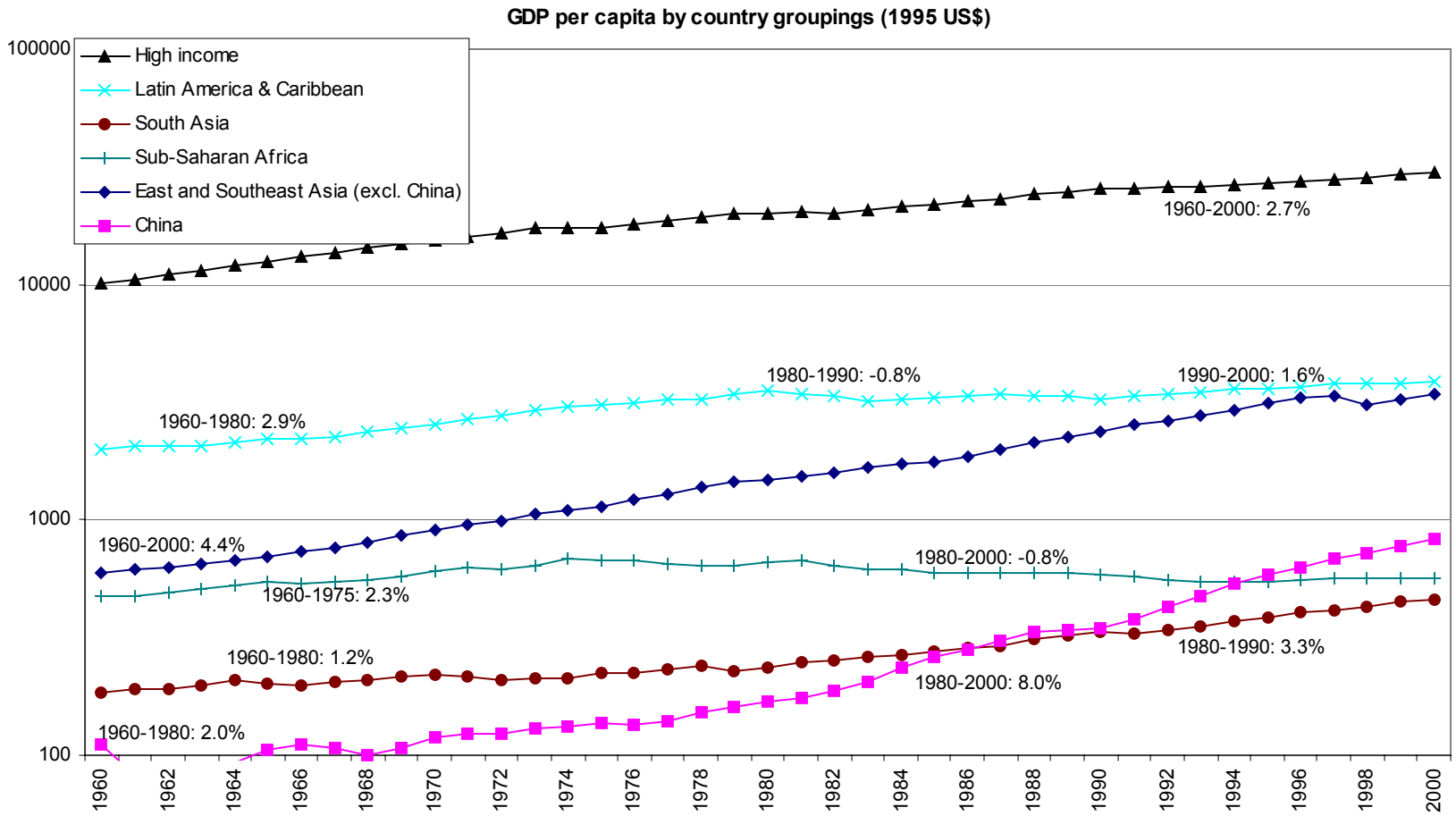

GDP per capita by country groupings (1995 US$)
High income
Latin America & Caribbean
South Asia
Sub-Saharan Africa
East and Southeast Asia (excl. China)
China
100000
10000
1000
100
1960-2000: 2.7%
1980-1990: -0.8%
1990-2000: 1.6%
1960-1980: 2.9%
1960-2000: 4.4%
1980-2000: -0.8%
1960-1975: 2.3%
1960-1980: 1.2%
1980-1990: 3.3%
1980-2000: 8.0%
1960-1980: 2.0%
1960
1962
1964
1966
1968
1970
1972
1974
1976
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000

Figure 1

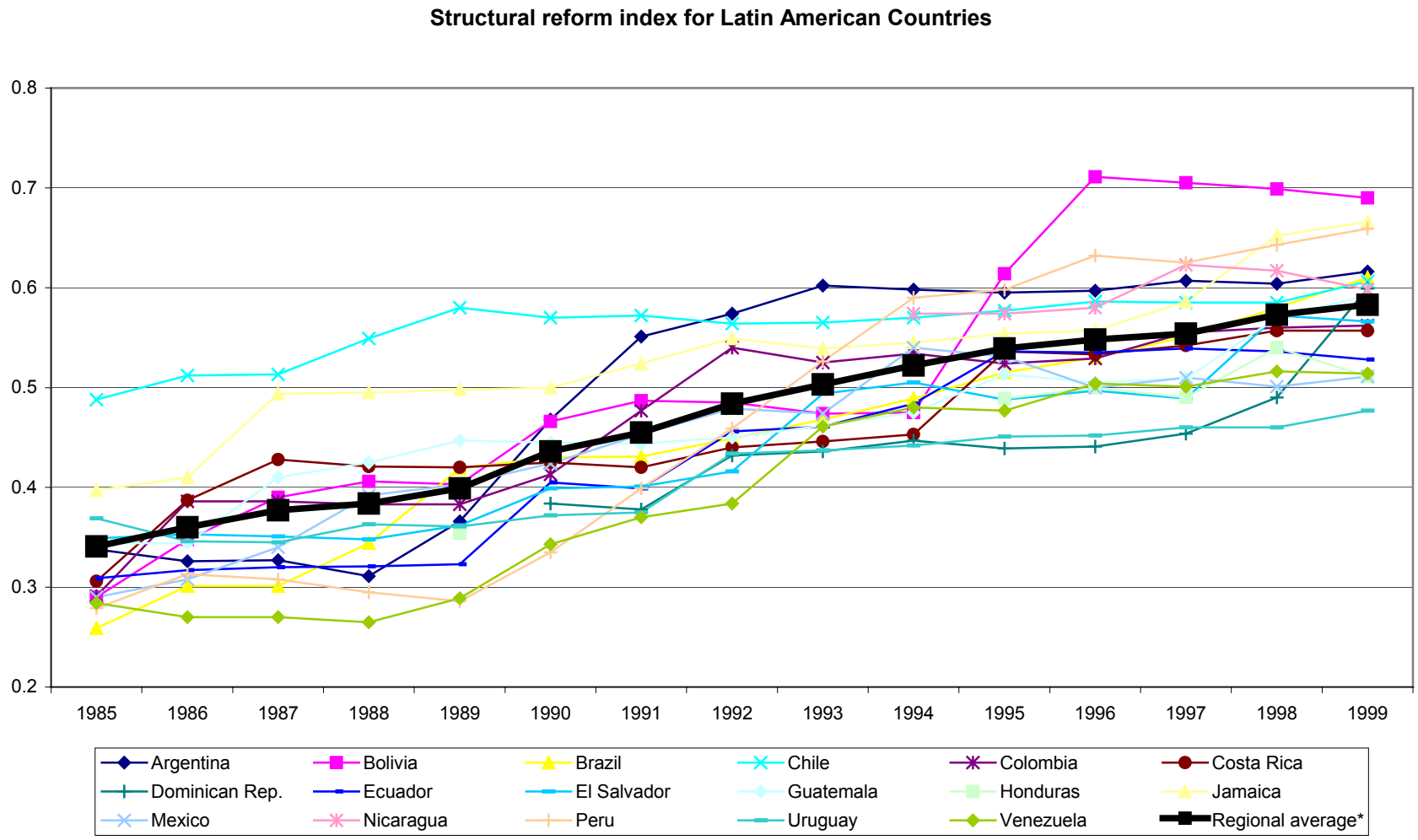


Figure 2

Source: Lora (2001a).

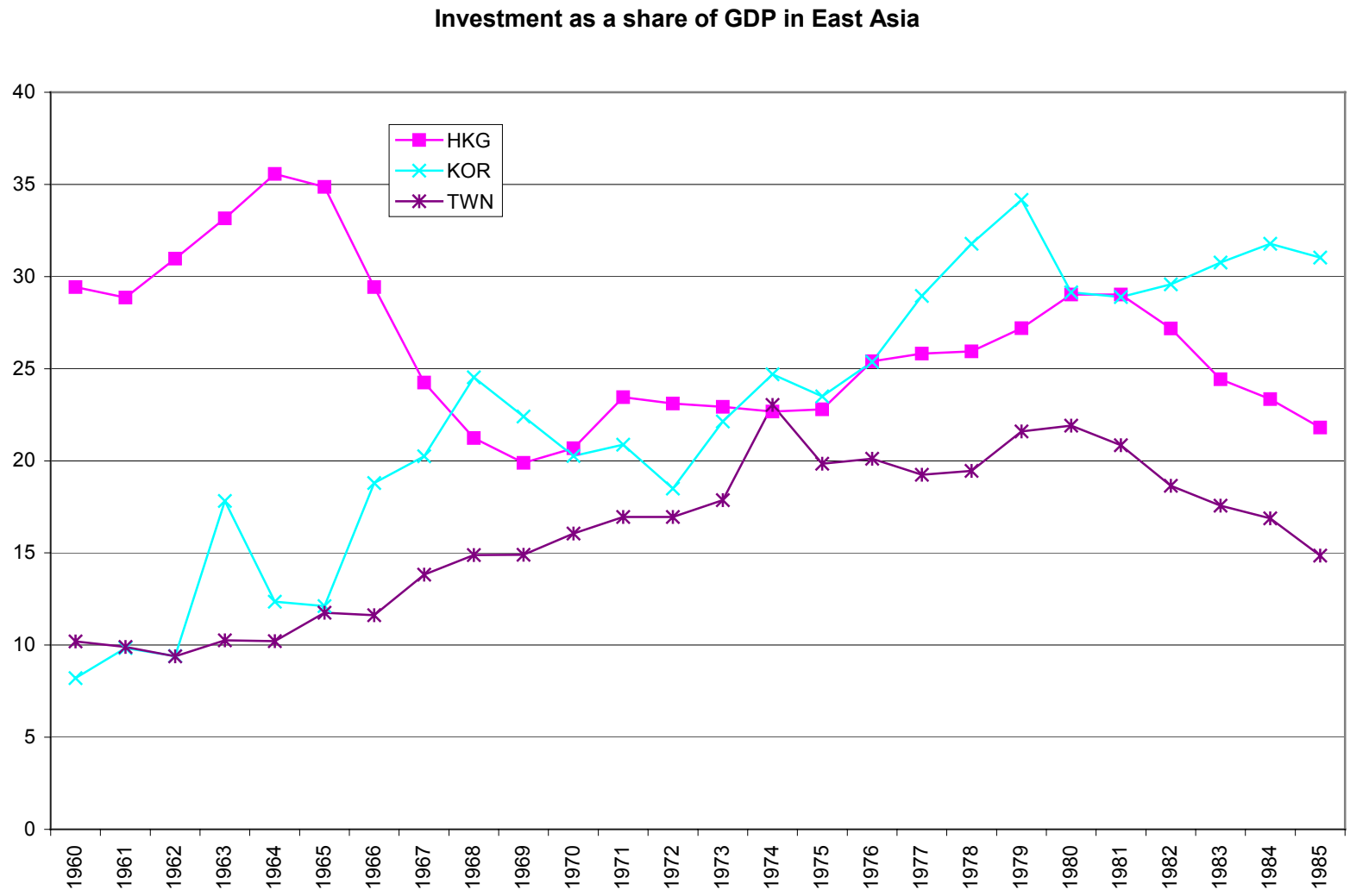

Investment as a share of GDP in East Asia
HKG
KOR
TWN
40
35
30
25
20
15
10
5
0
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985

Figure 3

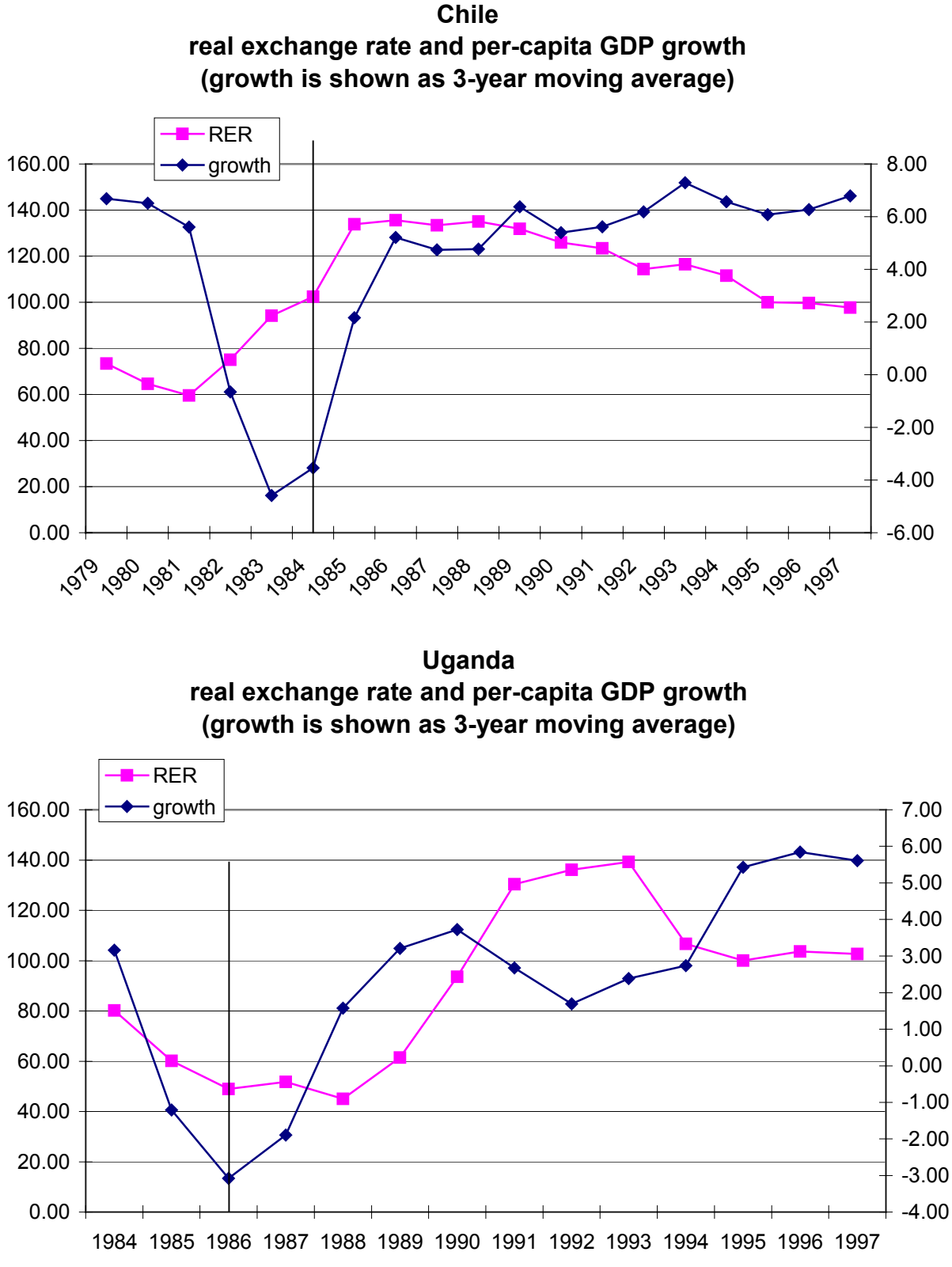


Figure 4: Real exchange rate and growth in Chile and Uganda

Chapter 15

NATIONAL POLICIES AND ECONOMIC GROWTH: A REAPPRAISAL*

WILLIAM EASTERLY

New York University Center for Global Development

Contents

* I am grateful from comments received at seminars at Pompeu Fabra, Boston University, and Brown University.

Handbook of Economic Growth, Volume 1A. Edited by Philippe Aghion and Steven N. Durlauf

DOI: 10.1016/S1574-0684(05)01015-4

Abstract

The new growth literature, using both endogenous growth and neoclassical models, has generated strong claims for the effect of national policies on economic growth. Empirical work on policies and growth has tended to confirm these claims. This paper casts doubt on this claim for strong effects of national policies, pointing out that such effects are inconsistent with several stylized facts and seem to depend on extreme observations in growth regressions. More modest effects of policy are consistent with theoretical models that feature substitutability between the formal and informal sector, have a large share for the informal sector, or stress technological change rather than factor accumulation.

Keywords

economic growth, macroeconomic policies, international trade, economic reform, economic development

JEL classification: O1, O4, E6, F4

An influential study by World Bank researchers Paul Collier and David Dollar (2001) finds that policy reform in developing countries would accelerate their growth and cut world poverty rates in half. They conclude that

> Poverty reduction – in the world or in a particular region or country – depends primarily on the quality of economic policy. Where we find in the developing world good environments for households and firms to save and invest, we generally observe poverty reduction.

I find the audacious claim that policy reform can cut world poverty in half a little daunting – even more so since Collier and Dollar base their results on an unpublished growth regression by me! (Like firearms, it is dangerous to leave growth regressions lying around.)

The International Monetary Fund (2000) also claims that "Where {sound macroeconomic} policies have been sustained, they have raised growth and reduced poverty." These claims are often held out as hope to economically troubled continents like Africa: "Policy action and foreign assistance ... will surely work together to build a continent that shows real gains in both development and income in the near future." Unfortunately, this claim was made in World Bank (1981) and the "real gains" in Africa have yet to arrive as of 2003.

Do the ambitious claims for the power of policy reform find support in the data? Are they consistent with theoretical views of how policy would affect growth?

The large literature on the determinants of economic growth, beginning with Romer (1986), has intensively studied national economic policies as key factors influencing long run growth. In this chapter, I take a look the state of this literature today, both theoretical and empirical. I do not claim to comprehensively survey the literature. I focus the chapter on the question of how strong is the case that national economic policies (by which I mean mainly macroeconomic and trade policies) have economically large effects on the growth rate of economies.

I am in the end skeptical that national policies have the large effects that the early growth literature claimed, or that the international agencies claim today. Although extremely bad policy can probably destroy any chance of growth, it does not follow that good macroeconomic or trade policy alone can create the conditions for high steady state growth.

Theoretical models that predict strong policy effects

The simplest theoretical model of endogenous growth is the *AK* mode of Rebelo (1991). Rebelo postulated that output could be proportional to a broad concept of capital (K) that included both physical and human capital:

$$Y = AK. \tag{1}$$

In principle, K could also include any kind of stock of knowledge, technology, or organizational technique that can be built up over time by sacrificing some of today's

consumption to accumulate such a stock. For example, technological knowledge could be accumulated by diverting some of today's output into lab equipment or other machines that help make new discoveries feasible. Or knowledge or human capital itself could be used to create further knowledge or human capital rather than producing today's output.[1] However, unlike many other endogenous growth models that explicitly address knowledge or technology [e.g. Aghion and Howitt (1998)], K is treated in this model as a purely private good – both excludable and rival. I will address below what happens when we relax this assumption.[2]

Constant returns to the factors that can be accumulated is also a key assumption in this model's prediction of a constant steady state rate of growth for given parameters and policies. This would rule out fixed costs in implementing a new technology, or increasing returns to accumulation at low levels of K, both of which feature in other growth models.

Since K is purely a private good, there is no role for government in this model. The market equilibrium yields the first best solution, and any government intervention in the form of taxes or price distortions must worsen welfare.

In this model, policies like tax rates have large effects on steady state growth. Consider first a tax (τ) on the purchase of investment goods (I). Consumption (C) is given by output less investment spending and taxes:

$$C = Y - (1+\tau)I. \tag{2}$$

Suppose the population size is constant and each (identical) household-dynasty maximizes welfare over an infinite horizon:

$$\max \int_0^\infty \mathrm{e}^{-\rho t} \frac{C^{1-\sigma}}{1-\sigma}\,\mathrm{d}t, \tag{3}$$

$$\dot{K} = I - \delta K. \tag{4}$$

[1] Rebelo (1991) showed that as long as the capital formation function itself has constant returns to accumulated factors, endogenous growth is possible even if final production has diminishing returns to capital.

[2] Since K in my models can always represent either technology or factor accumulation, I do not address the hot debate on how much factor accumulation matters for growth. On education, Benhabib and Spiegel (1994) and Pritchett (1997) show that cross-country data on economic growth rates show that increases in human capital resulting from improvements in the educational attainment of the work force have *not* positively affected the growth rate of output per worker. It may be that, on average, education does not effectively provide useful skills to workers engaged in activities that generate social returns. There is disagreement, however, Krueger and Lindahl (1999) argue that measurement error accounts for the lack of a relationship between growth per capita and human capital accumulation. Hanushek and Kimko (2000) find that the quality of education is very strongly linked with economic growth. However, Klenow (1998) demonstrates that models that highlight the role of ideas and productivity growth do a much better job of matching the data than models that focus on the accumulation of human capital. More work is clearly needed on the relationship between education and economic development. On physical capital accumulation, there is the debate between the "neoclassical" school stressing factor accumulation [Mankiw, Romer and Weil (1992), Mankiw (1995), Young (1995)] and the school stressing technology or the residual [Klenow and Rodriguez-Clare (1997a, 1997b), Hall and Jones (1999), Easterly and Levine (2001)].

Then the consumer–producer would invest at a rate that results in steady-state growth of

$$\frac{\dot{C}}{C} = \frac{(A/(1+\tau)) - \delta - \rho}{\sigma}. \tag{5}$$

Here policy has large effects on steady state growth. If $A = 0.15$ and $\sigma = 1$, then an increase from a tax rate of 0 to one of 30% would lower growth by 3.5 percentage points. Such a policy pursued over 30 years would leave income at the end 65 percent lower than it would have been in the absence of a tax. This is a strong claim for the effects of policy on economic development! It offers a possible explanation for the poverty of a poor nation – bad government policies (high τ) – which can be remedied easily enough by changing to good policies (low τ). It is clear why this has been a seductive theory for aid agencies and policymakers that seek to promote economic development.

The effects on accumulation are even more dramatic. Solving for the broad concept of investment that includes physical capital, human capital, technology, and knowledge accumulation, we get:

$$\frac{I}{Y} = \frac{(A/(1+\tau)) - \delta(1-\sigma) - \rho}{\sigma A}. \tag{6}$$

The effect of taxation on investment does not depend on A. If $\sigma = 1$, the derivative of I/Y with respect to the tax factor $1/(1+\tau)$ is unity. An increase of the tax rate from 0 to 30 percent would reduce investment by 23 percentage points of GDP!

Before examining this claim in more detail, note that the tax rate on investment goods does not have to be an explicit tax on capital goods. First of all, there is an equivalent income tax that would have had the same effect on growth (given by $t = 1 - 1/(1+\tau)$), so policies here could be any government action that diverts income away from the original investor in production. (Note using the result above, that every one percentage point increase in the income tax rate reduces investment by one percentage point of GDP.) Second, note that this result applies to the *marginal* effective tax rate on investment goods or income. While movements from 0 to 30 percent would be dramatic for average tax rates, a movement of 30 percentage points in marginal effective tax rates could easily come from a tax reform. Second, the tax on capital goods could stand for any policy that alters the price of investment goods relative to consumption.[3] For example, suppose that a populist government controls output prices for consumers but the investor must buy goods for investment on the black market. Then the premium of the black market price over the official price would act much like a tax on investment goods. If the one good in this model is tradeable, then the black market premium on foreign

[3] Chari, McGrattan and Kehoe (1996) and McGrattan and Schmitz (1998) present models and empirical work emphasizing the measured high relative price of capital goods as a policy factor inhibiting economic development. Hsieh and Klenow (2003) have an alternative story that stresses high capital prices and low income as the joint outcome of a technological disadvantage in producing tradeable goods (including capital goods) in poor countries.

exchange might be a good proxy for the wedge between official output prices and black market investment good prices (assuming that consumer goods can be imported at the official exchange rate, or at least that official output prices are controlled as if they could be). If we suppose that the purchaser of investment goods must hold cash in advance of a purchase of investment goods, then inflation would be indirectly be a tax on investment goods. One could also get similar results with institutional variables – a probability of expropriation of part or all of the capital good by the government or government officials demanding a bribe every time a new unit of capital is installed would act much like a tax on investment.

The claims for large policy effects become even stronger in growth models with increasing returns to capital and externalities. Suppose that there is a group of large but fixed size where the capital held by each member of the group has non-pecuniary externalities for the rest of the group. For example, a high human capital individual in a residential neighborhood might benefit the rest of the neighborhood with whom she socially interacts. The knowledge and connections that this individual brings might raise the productive potential of others (this is loosely what is called "social capital" in the literature). If this is true for all social interactions in the neighborhood, and these interactions are identical, costless, and exogenous for all members, then there will be a spillover from the average human capital of the neighborhood to each inhabitant of the neighborhood. The production function for an individual member would look like this:

$$y = Ak^{\alpha}\bar{k}^{\beta}. \tag{7}$$

One can think of other similar examples of spillovers. If k includes knowledge or technology, it is plausible that these goods are non-rival and partially non-excludable. For example, firms may benefit by example from new technology installed by other firms in the same trade. People in almost every human activity engage in "shop talk" that is incomprehensible to outsiders, but which apparently conveys productive knowledge to those involved in the activity.[4]

Assuming the same maximization problem as above (Equations (2) through (4)), then the individual will invest in k taking everyone else's investment as given (because the group is too large for her to influence its average). The optimal path of consumption is now given by

$$\frac{\dot{C}}{C} = \frac{(A\alpha k^{\alpha-1}\bar{k}^{\beta}/(1+\tau)) - \delta - \rho}{\sigma}. \tag{8}$$

However, since all members of the group are assumed to be identical, then $k = \bar{k}$ ex-post, and the growth rate for each individual will be

$$\frac{\dot{C}}{C} = \frac{(A\alpha \bar{k}^{\alpha+\beta-1}/(1+\tau)) - \delta - \rho}{\sigma}. \tag{9}$$

[4] The emphasis on the special properties of knowledge and technology was highlighted by Romer (1995) and Aghion and Howitt (1998). The idea of social capital has been stressed by authors such as Putnam (1993, 2000), Glaeser (2000), Narayan and Pritchett (1997), Woolcock and Narayan (2000).

There are multiple equilibria if $\alpha + \beta - 1 > 0$, i.e. if both the original importance of broad capital to production is large plus there are strong spillovers. If we have the special case of $\alpha+\beta = 1$, then we are back to the AK model, albeit one with suboptimal market outcomes because of the externality. If $\alpha+\beta-1 < 0$, then the model will feature similar prediction as the neoclassical model with a high capital share (discussed below).

In the multiple equilibria case, the return to capital increases the more initial capital there is, the opposite of the usual diminishing returns to capital. Figure 1 illustrates the possible outcomes. If the tax rate is low, the after tax rate of return to capital is the upper upward-sloping line. Any initial capital stock to the left of point A (where the after tax return is less than $\delta + \rho$) will go into a vicious circle of negative growth of consumption and decumulation of capital. Any point to the right of A (such as B) will go into a virtuous circle of positive and accelerating growth of consumption and positive capital accumulation.[5] Now suppose that tax rates are increased, shifting the rate of return to the lower upward-sloping line in Figure 1. Now any point to the left of C will go into a vicious circle of decline. An economy with capital stock B, which was in the expanding region under low taxes, is now in the declining region under high taxes. A policy shift now has an even more dramatic impact on national prosperity – it could spell the difference between subsistence consumption (say Mali) and industrialization (say Singapore). Policy spells the difference in the long run between per capita income of $300 and $30,000 – rather a dramatic effect. As in all multiple equilibria models, initial conditions matter and small things (like policy) can have large consequences. If the first endogenous growth model was seductive to policymakers, this is even more so – one government official at the stroke of a pen could change a nation's prospects from destitution to prosperity.

This increasing returns model is much like poverty trap models like those of Azariadis and Drazen (1990), Becker, Murphy and Tamura (1990), Kremer (1993), and Murphy, Shleifer and Vishny (1989). It is also consistent with models of in-group ethnic and neighborhood externalities [Borjas (1992, 1995, 1999), Benabou (1993, 1996)] and geographic externalities [Krugman (1991, 1995, 1998), Fujita, Krugman and Venables (1999)]. Ades and Glaeser (1999) present evidence for increasing returns in closed economies.

A story like that told in Figure 1 would also predict instability of growth rates if an economy is in the middle region B and is subject to continuous fluctuations in policies. The economy would keep shifting from positive to negative growth and back again as policies change. This is a possible story for some of the spectacular reversals in output growth that we have seen in countries like Cote d'Ivoire, Jamaica, Guyana, and Nigeria (see Figure 2).

It is often assumed that these strong claims for policy effects on growth are only a feature of endogenous growth models. However, the other innovation in the growth

[5] The feature of ever accelerating growth in this model leads to nonsensical predictions in the long run – the model would have to be modified at higher incomes with some feature that puts a ceiling on the rate of return to capital.

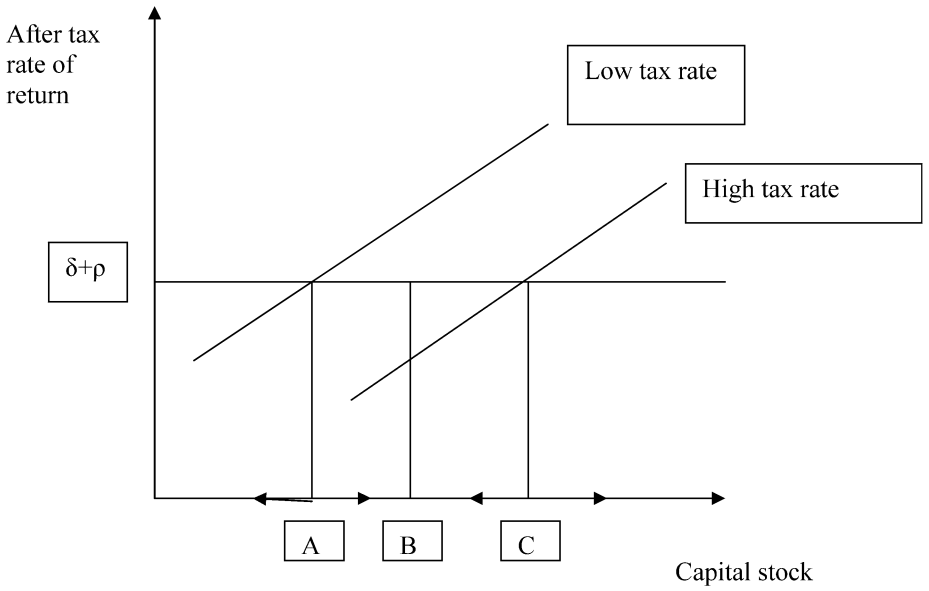


Figure 1. Multiple equilibria with increasing returns to capital, alternative tax regimes.

literature of the last two decades has been to put a much higher weight on capital even in the neoclassical exogenous growth model. Again, the justification is that capital is a broader concept than just physical equipment and buildings. It should include at least human capital, if not the more technology and knowledge forms of capital discussed above. Attributing part of the labor income in the national accounts to human capital, this would raise the share of capital in output from around 1/3 (if the only form of capital was physical) to something like 2/3.[6] The high capital share is also necessary to avoid counterfactual predictions about very high returns to capital in capital-scarce countries, and the same in the initial years of a transition from capital-scarcity to capital-abundance.

The neoclassical production function with labor-augmenting technological change is:

$$Y = K^{\alpha}(AL)^{1-\alpha}. \tag{10}$$

In per capita terms, we have:

$$y = k^{\alpha}A^{1-\alpha}. \tag{11}$$

The consumer–producer's maximization problem is the same as before, using Equations (2) through (4). Technological progress (the percent growth in A) is assumed to

[6] Mankiw, Romer and Weil (1992) and Mankiw (1995).

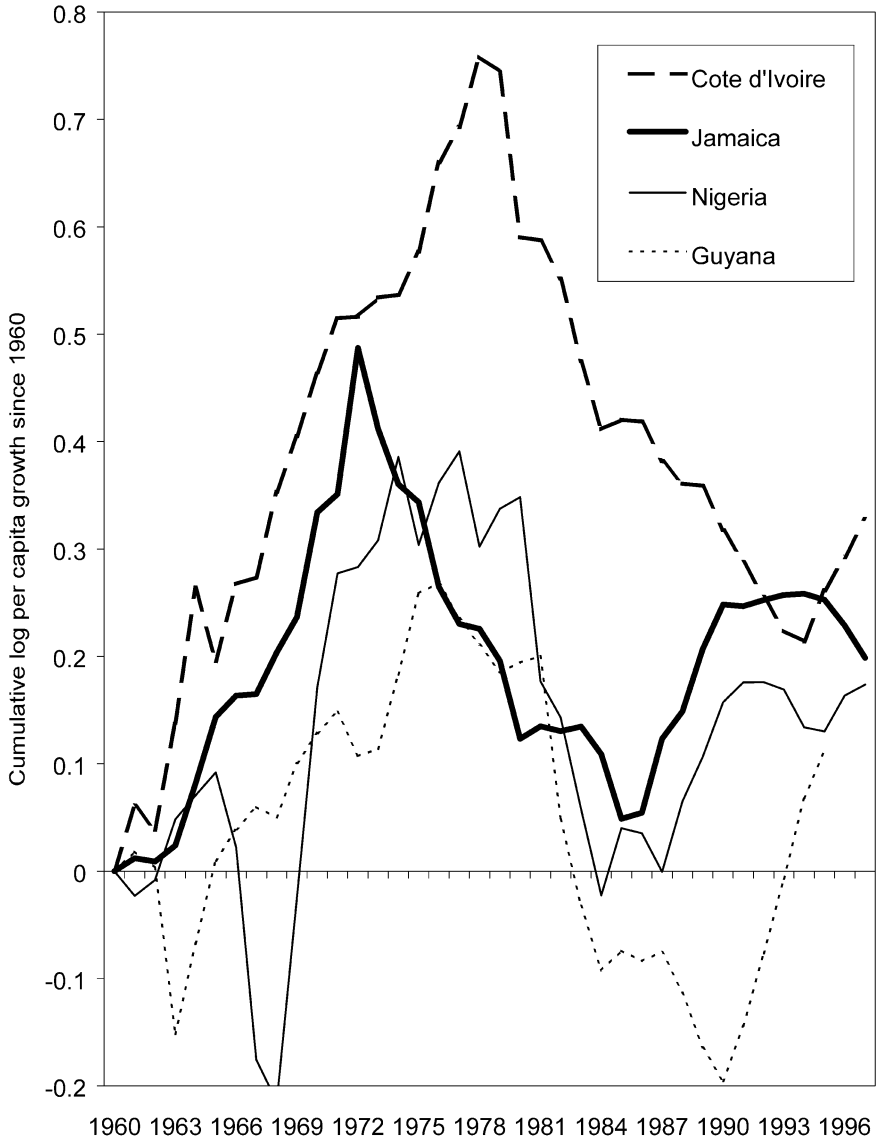


Figure 2. Examples of variable per capita income over time.

take place at an exogenous rate x. As is well known, accumulation of physical and human capital cannot sustain growth in the long run in the absence of technological progress. Since policy affects the outcome only through the incentive to accumulate capital, it follows that policy by itself cannot foster sustained growth in this model.

With growth in A of x, the long-run steady state will have per capita output y, capital per worker k, and per capita consumption all growing at the same (exogenous) rate x. The tax rate on capital goods has no effect on the steady-state growth rate. However, policy does have potentially large effects on the level of per capita income. To see this, it is convenient to write both capital per worker and per capita income relative to the technological level A. The optimal growth of per capita consumption is now:

$$\frac{\dot{C}}{C} = \frac{(\alpha(k/A)^{\alpha-1}/(1+\tau)) - \delta - \rho}{\sigma}. \tag{12}$$

Since (12) must equal x in steady state, an increase in the tax rate τ must always be offset by a decrease in the relative capital stock (raising the pre-tax rate of return to capital because of diminishing returns, i.e. because $\alpha < 1$). Setting (12) equal to x determines the k/A ratio in the steady state, which in turn gives the following for per capita income relative to technology:

$$\frac{y}{A} = \left[\frac{\alpha}{(1+\tau)(\sigma x + \delta + \rho)}\right]^{\frac{\alpha}{1-\alpha}}. \tag{13}$$

A high tax on investment inhibits capital accumulation and thus lowers the level of income relative to the technology level. High taxes are still a possible explanation of relative poverty in the neoclassical model. With a capital share of 2/3 (including both human and physical capital), a tax rate decrease from 50 percent to zero raises income by a factor of $(1.5)^2$, or 2.25 times. If the capital share were 0.8 (as writers like Barro and Mankiw have suggested), then the tax reduction would raise income by a factor of $(1.5)^4$ or 5 times.

Although there is no effect of the tax change on steady state growth, there will be a dramatic change in growth in the transition from one policy regime to another. There is one unique saddle path to the new steady state; consumption will jump to that saddle path after the change in policy (in a world of perfect certainty of course). To solve for the transition involves solving for the saddle-path of consumption in transition to the new steady state. Figure 3 shows a simulation of a decrease in the tax rate on investment from 50 percent to zero, with the following parameter values:

$\alpha = 0.6666$,
$\delta = 0.07$,
$\rho = 0.05$,
$\sigma = 0.9$,
$x = 0.02$.

For comparison, I also show a simulation of an endogenous growth rate model with $A = 0.138$, which gives the same 2 percent per capita growth rate at zero tax as the exogenous growth neoclassical model. Both models show dramatic growth rate effects after the policy change, still large after 20 to 30 years. It is only in the very long run that the neoclassical growth effect wears off with diminishing returns. Investment rates would show similar jumps after the policy change as growth rates.

aghion v.2005/08/19 Prn:23/08/2005; 8:39 F:aghion1015.tex; VTEX/JK p. 11
aid: 1015 pii: S1574-0684(05)01015-4 docsubty: REV

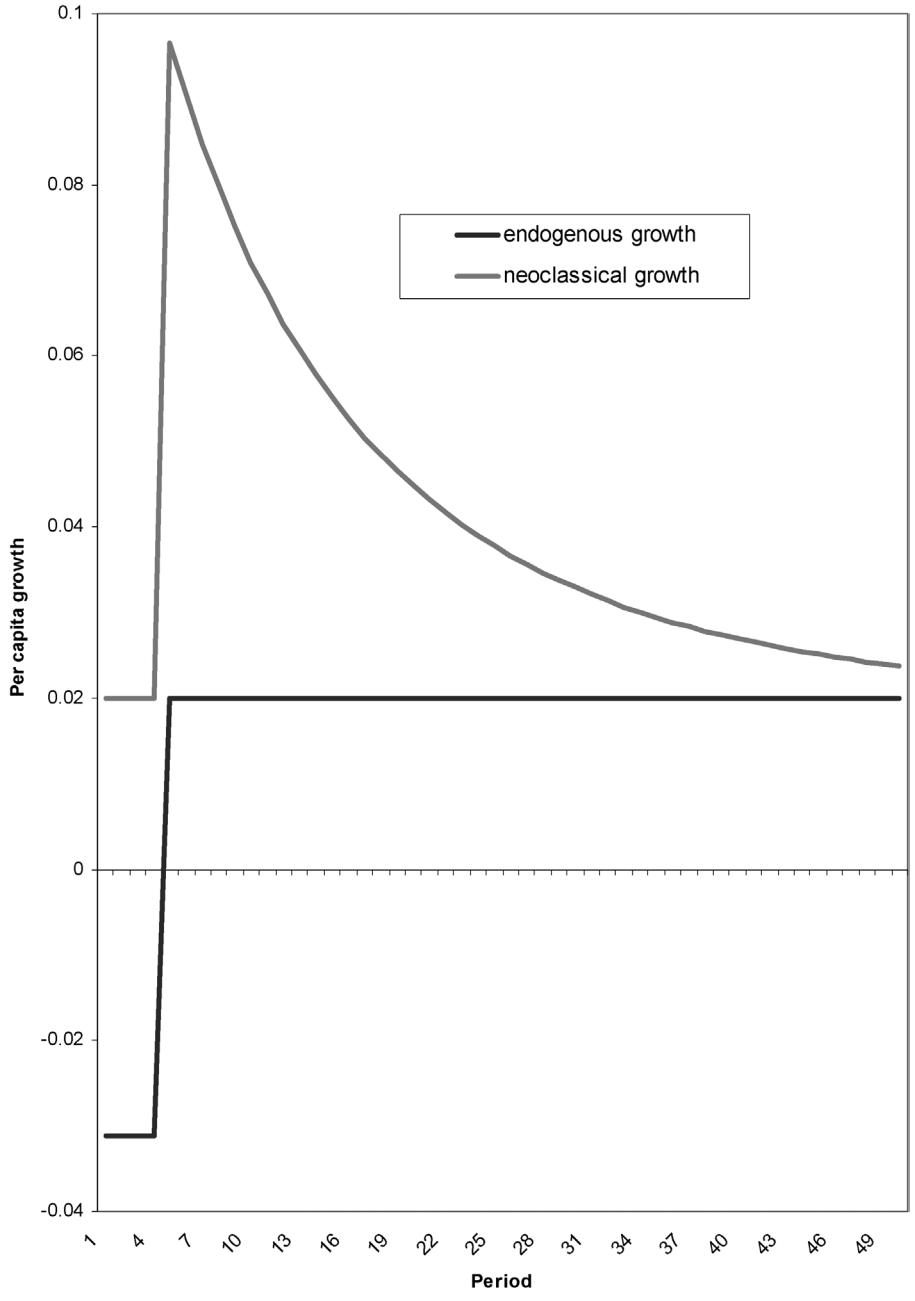


Figure 3. Endogenous growth and neoclassical growth with a reduction of tax rate on investment from 50 percent to zero.

What is different for the purposes of empirical work is that the predicted difference in growth rates in the endogenous growth model before and after the tax decrease could equally apply to cross-section differences in growth between high-tax and low-tax countries. In the neoclassical model, the predicted effect of policy change on growth is only for a cross-time effect within countries. However, this difference has been handled in practice by testing the effect of current policies on growth, controlling for initial income. Initial income can be thought of as representing policy regimes prior to the period under study. If current policy predicts a higher steady state level of income than initial income, then the transitional dynamics like those shown in Figure 3 will be set in motion. The neoclassical model would predict instability of growth rates over time if frequent policy changes shift the steady state level above or below the current income level, which is ironically similar to the increasing returns prediction of growth rate instability.

One big difference between the three models is that the neoclassical model predicts falling growth and investment after the initial policy-induced increase in growth, the increasing returns to capital model predict rising growth and investment afterwards, while the constant returns to capital model predict constant growth. I will examine some case studies of major policy reforms below to see which of these predictions appears to hold.

All of the three models predict large growth effects of policy changes. I will examine below the evidence for or against these claims, but here I will note how much these bold predictions are different from many other fields of economics, as well as from the pre-1986 growth literature. The literature on tax policy, for example, thinks that it is a big deal to identify a benefit of 0.1 percent of GDP from a major tax reform that lowers distortions. The notion that economic development of a whole society can be achieved a few stroke-of-the pen policy reforms seems simplistic in retrospect. If this is so, why haven't more countries successfully developed? Are large policy effects on growth an inevitable feature of new growth models?

Models that predict small policy effects on growth

To begin to understand some of the factors that might mitigate the large effects of policy on growth, suppose that there output is a function of two types of capital, only one of which can be taxed. For example, suppose that the first type of capital (K_1) is formal sector capital that must be transacted on markets in the open, while the second type of capital (K_2) is informal sector capital that can be accumulated away from the prying eyes of the tax man.

$$Y = A\left(\alpha K_1^{\gamma} + (1-\alpha)K_2^{\gamma}\right)^{\frac{1}{\gamma}}, \tag{14}$$

$$C = Y - (1+\tau)I_1 - I_2, \tag{15}$$

$$\dot{K}_1 = I_1 - \delta K_1, \tag{16}$$

$$\dot{K}_2 = I_2 - \delta K_2, \tag{17}$$

$$\frac{\dot{C}}{C} = \frac{(A\alpha[\alpha + (1-\alpha)(K_2/K_1)^{\gamma}]^{\frac{1-\gamma}{\gamma}}/(1+\tau)) - \delta - \rho}{\sigma}. \tag{18}$$

If these two capital goods are close to perfect substitutes, then the effects of taxes on growth go towards zero. Figure 4 shows the relationship between growth and tax rates at extreme values of γ. With γ close to 1 (close to perfect substitutability), there is only a minor effect of taxes and it is bounded from below no matter how high the tax rate. This is because with the elasticity of substitution greater than one, formal sector capital is not essential to production. The worst that high tax rates can do is drive formal capital use down to zero (which has only a small effect if the capital goods are close to perfect substitutes). After that, increases in tax rates have no further effect (explaining the flat segment of the curve in Figure 4). The effects of tax rates on growth continue to be strong if the elasticity of substitution between the two goods is less than one (the $\gamma = -1$ line in Figure 4).

The other parameter that plays an important role in how damaging are tax rates is the share (α) of formal sector capital (or more specifically, the share of the capital that is actually subject to taxation). Figure 5 shows how different are the effects of taxing investment in this factor when its share (α) is 0.1 compared to when its share is 0.8 (assuming an elasticity of substitution of unity). Of course, lowering the share of taxable capital would also limit the power of taxation in the neoclassical model.

Another factor that mitigates the effects of policies on growth is that many policies distort relative prices amongst different sectors or different types of goods, rather than penalizing all capital goods. With a distortion of relative prices, some capital goods are more expensive but others are cheaper. For example, with a black market premium on foreign exchange, those who receive licenses to import goods at the official exchange rate receive a subsidy, while those who must pay the black market rate for inputs pay an implicit tax.[7] Unanticipated high inflation is a tax on creditors but a subsidy to debtors. An overvalued real exchange rate penalizes producers of tradeables but subsidizes producers of nontradeables. Trade protection taxes imports but subsidizes production for the domestic market. The rate of subsidy is clearly related to the rate of taxation. One way to pin it down is to specify that the revenues from the tax on the first type of capital must just cover the subsidy expenditures on the second type of capital.

Here are the equations I have in mind. I revert to Cobb–Douglas for simplicity:

$$Y = AK_1^{\alpha}K_2^{1-\alpha}, \tag{19}$$

$$C = Y - (1+\tau)I_1 - (1-s)I_2. \tag{20}$$

(16) and (17) still represent the capital accumulation equations, and the consumer-producer maximizes (3) taking τ and s as given. Ex-post, the government must balance

[7] If black markets function efficiently, the opportunity cost of inputs is their black market value even for those who receive them at the subsidized price. However, the recipient of inputs at the official exchange rate still receives a subsidy per unit of input use.

aghion v.2005/08/19 Prn:23/08/2005; 8:39 F:aghion1015.tex; VTEX/JK p. 14
aid: 1015 pii: S1574-0684(05)01015-4 docsubty: REV

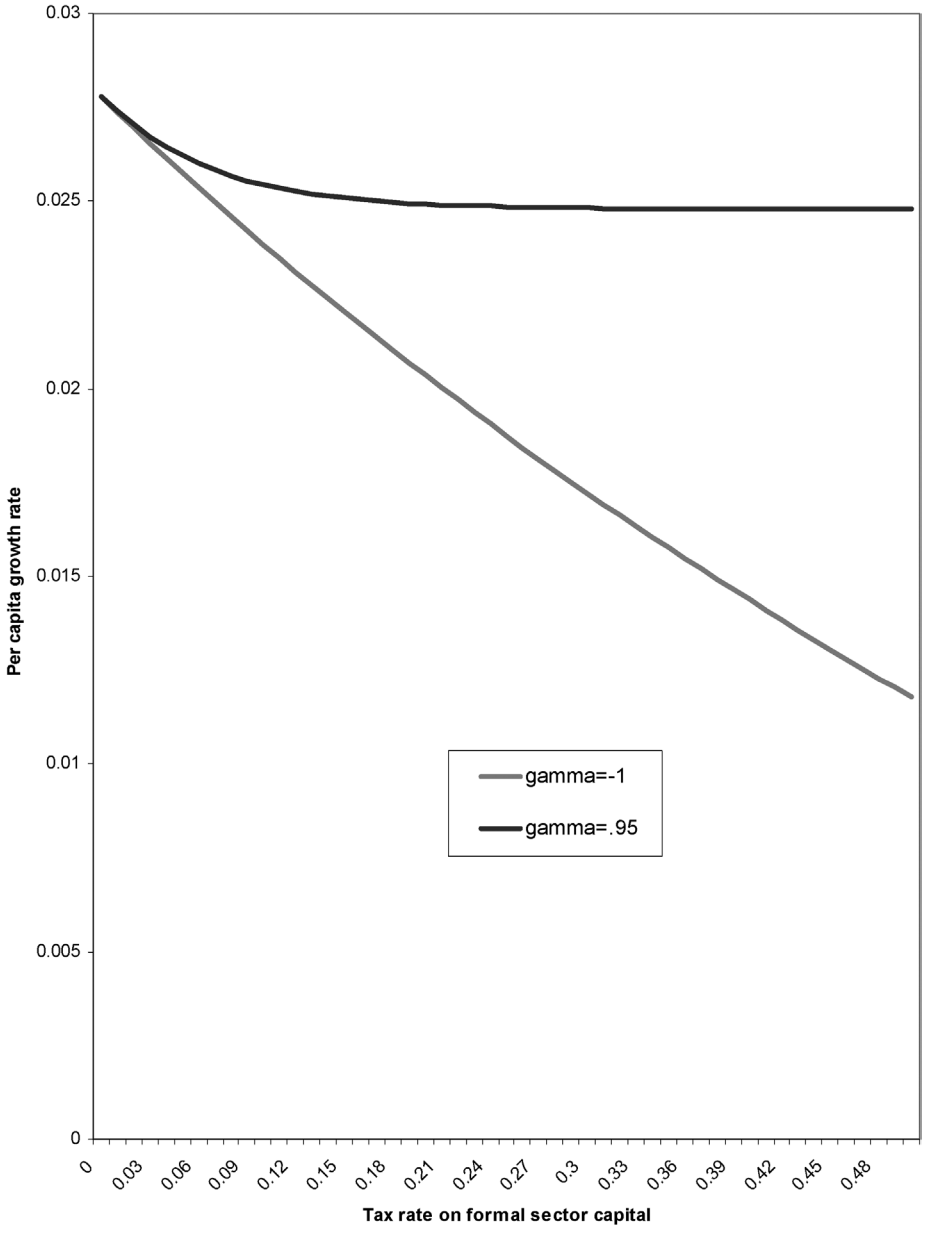


Figure 4. Growth rates with different assumptions about elasticity of substitution between capital good types.

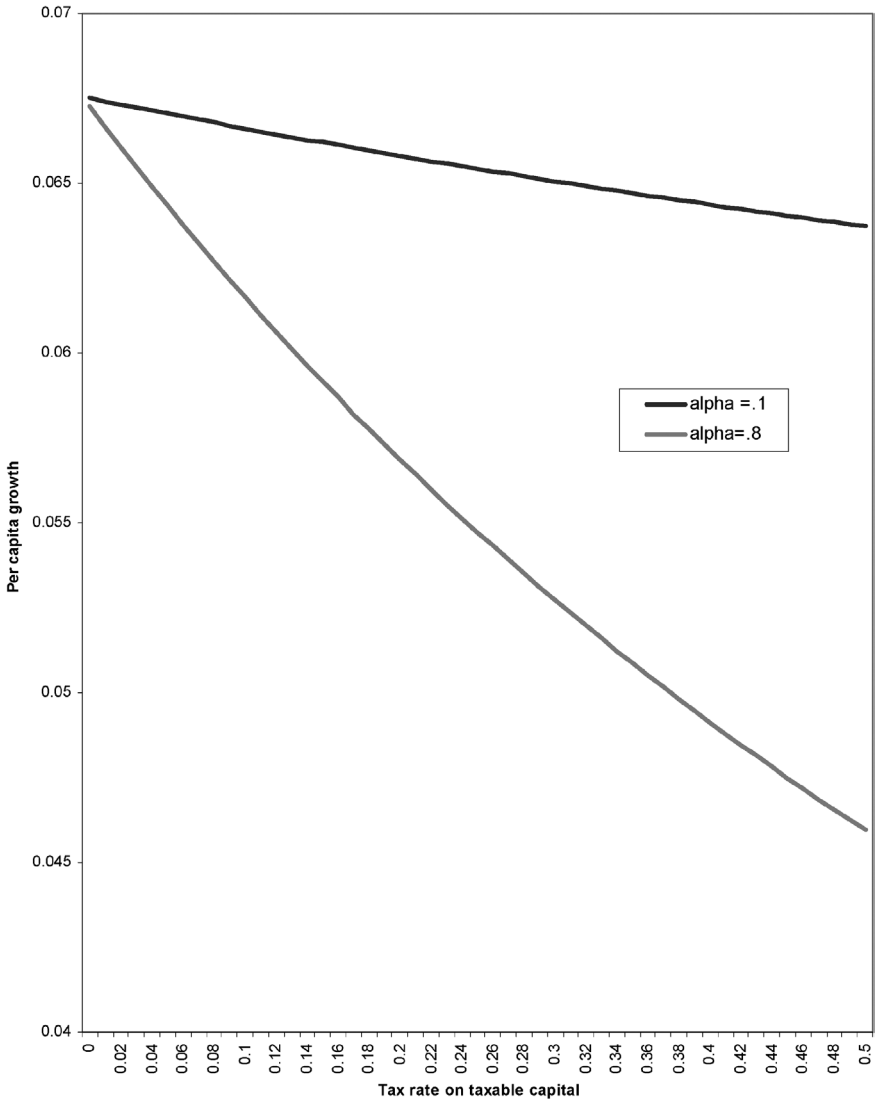


Figure 5. Tax rates and growth with different shares of taxable capital.

its budget so:

$$\tau I_1 = s I_2. \tag{21}$$

Because of the neat properties of Cobb–Douglas, the solution of the optimal capital ratio as a function of the subsidy rate (after taking into account the fiscal relation-

aghion v.2005/08/19 Prn:23/08/2005; 8:39 F:aghion1015.tex; VTEX/JK p. 16
aid: 1015 pii: S1574-0684(05)01015-4 docsubty: REV

ship (21) between tax rates and subsidy rates) is very simple:

$$\frac{K_2}{K_1} = \frac{1-\alpha}{\alpha - s}. \tag{22}$$

The growth rate will display offsetting effects of the subsidy-cum-tax rate – on the one hand, it distorts the allocation of capital away from K_1 to K_2, lowering the pre-subsidy marginal product of K_2, while on the other hand, it of course subsidizes the rate of return to K_2.

$$\frac{\dot{C}}{C} = \frac{(A(1-\alpha)((\alpha - s)/(1-\alpha))^{\alpha}/(1-s)) - \delta - \rho}{\sigma}. \tag{23}$$

One can show that if (21) (the balanced budget requirement) is imposed, it is impossible for this kind of tax-cum-subsidy scheme to raise the rate of growth.[8] The tax-cum-subsidy will imply an efficiency loss from the distortion of resource allocation, and this efficiency loss will have a negative growth effect if all types of capital can be accumulated. However, the relationship between the distortion and the growth rate is highly nonlinear. As is well known in the literature on relative price distortions, the cost of the distortion increases more than proportionately with the size of the distortion.[9] In the traditional literature on "Harberger triangles", this was an output loss. In an endogenous growth model where all inputs can be accumulated, the distortion between relative prices of the inputs induces a reduction in growth. A small distortion introduces only a small wedge in between marginal products of the two inputs and does not cause a huge growth loss. Eventually, however, the distortion forces far too much accumulation of one type of capital relative to the other, severely lowering the marginal product of the excessive capital good due to diminishing returns. An increasing rate of subsidy also requires a more than one for one increase in the tax rate, as the tax base is shrinking with increased taxes while the capital goods being subsidized are increasing. The nonlinear relationship is shown in Figure 6. Note that distortions do not have much effect on growth at all up to subsidy rates of about 0.2 and then have increasingly catastrophic consequences after about 0.4.

There are other factors that mitigate the effects of policy on growth that I do not explicitly model here. One is policy uncertainty. The announcement of a new policy may not be credible, perhaps because high political opposition to it may imply a high probability of subsequent reversal. Many developing countries have a history of frequent reversals of incipient policy reforms, which makes any future reform less believable. For example, Argentina has been a chronic high inflation country for nearly half a century. Frequent stabilization attempts have subsequently come unwound; the fiasco of the Convertibility Plan in 2001 is only the latest example. In terms of the model above, the

[8] This applies to CES production functions more generally [see Easterly (1993) for a proof].

[9] One recent growth model emphasizing this nonlinearity is Gylfason (1999), where the cost e of a distortion c is amusingly expressed as $e = mc^2$.

aghion v.2005/08/19 Prn:23/08/2005; 8:39 F:aghion1015.tex; VTEX/JK p. 17
aid: 1015 pii: S1574-0684(05)01015-4 docsubty: REV

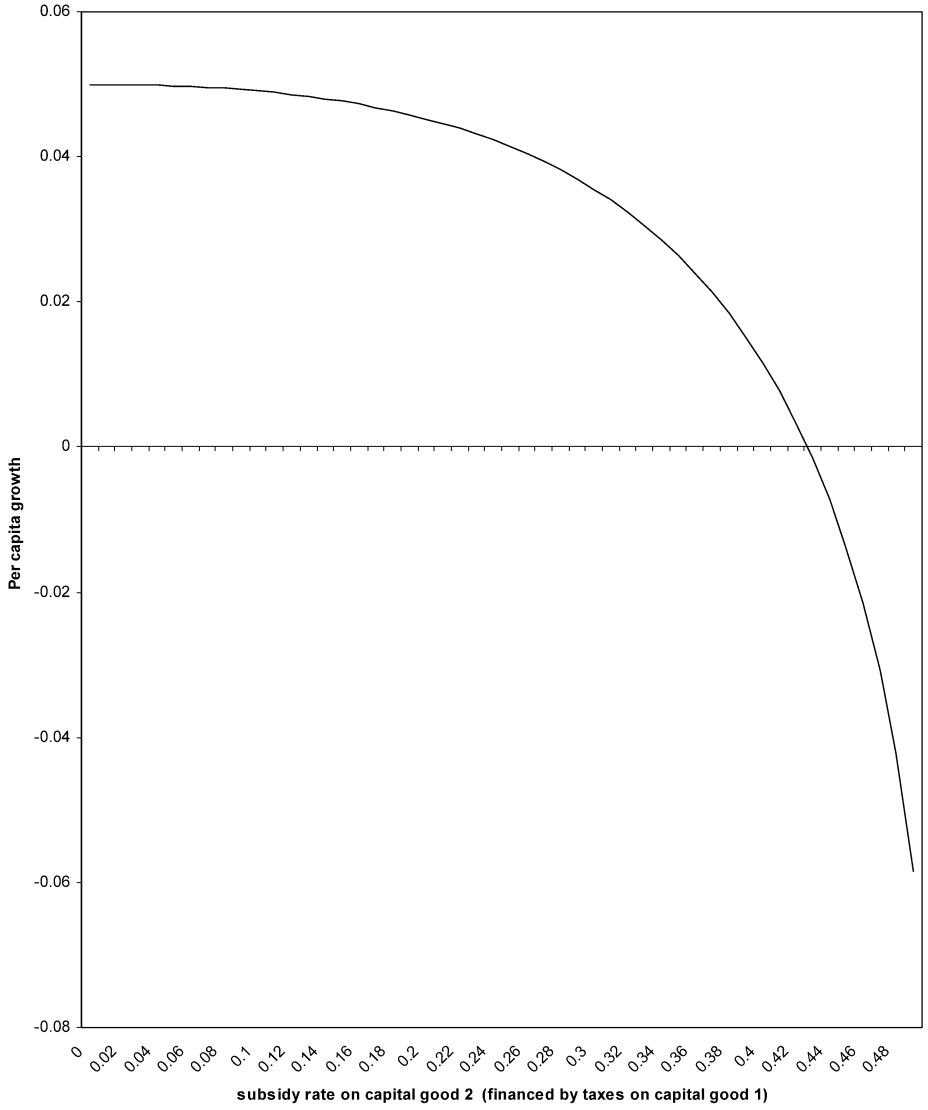


Figure 6. Growth rate and subsidy rate financed by taxes.

certainty equivalent of the after-tax return on capital may not increase much even after an announcement that taxes will be cut.

There is also the possibility that policies whose main purpose was to create rents for political patronage will be replaced with other policies that create new rents. For example, if the black market premium is abolished, the holders of import licenses at the

official exchange rate may seek new sources of income (for example, appointment as customs inspectors, where they can take bribes). There may be a law of conservation of political rents, akin to the second law of thermodynamics, if the factors inducing political rent seeking do not change.

Poor countries may be so close to subsistence consumption that they may not be able to take advantage of policy changes. Rebelo (1991) and Easterly (1994) show intertemporal utility functions with Stone–Geary preferences, in which consumers derive utility from consumption only above a certain floor of subsistence. This model predicts a very low intertemporal elasticity of substitution at levels of consumption close to subsistence. Intuitively, consumers close to subsistence have a limited ability to postpone consumption in order to take advantage of higher returns to saving. This model predicts a slow acceleration of growth even after a favorable policy change, as consumption must first rise well above subsistence.

Most importantly, policies may be offset or reinforced by more important factors that affect the growth and income. Achieving high output returns from a given set of inputs involves an incredibly complex set of institutions (such as enforcement of contracts and property rights), social norms, efficient sorting and matching of people and other inputs, advanced technological knowledge, full information on both sides of all transactions, low transaction costs, resolution of principal-agent problems, positive non-zero-sum game theoretic interactions among agents, resolution of public good problems, and so on. The development of institutions and social and political structures that address these issues successfully (from the standpoint of material production) is probably a long historical process.

The above models have a pale shadow of all this complexity in the parameter A. Note that the lower is A, the lower is the derivative of growth (or income in the neoclassical model) with respect to the policy parameter τ. Many authors have argued that differences in A explain a large part of income differences between countries [Hall and Jones (1999), Klenow and Rodriguez-Clare (1997a, 1997b), Easterly and Levine (2001)]. If a poor country is poor because of low A, then a change in policies may not do much to raise income or growth. Exogenous variation in A may also affect the political economy of policy – a high A country would be less likely to tolerate the costs of destructive policies, while bad policy may be tolerated in a low A country because it may not make much difference. Of course policy itself could influence A. However, if A really depends on all the complexities listed above, then the kind of macroeconomic policies I am considering in this paper may not have much effect on A.

Empirics

The literature tracing effects of economic policies on growth is abundant. I do not attempt to summarize it here, noting the summaries in Sala-i-Martin (2000), Temple (1997), Kenny (2001), and Easterly and Levine (2001). Some authors focus on openness to international trade [Frankel and Romer (1999)], others on fiscal policy [Easterly

aghion v.2005/08/19 Prn:23/08/2005; 8:39 F:aghion1015.tex; VTEX/JK p. 19
aid: 1015 pii: S1574-0684(05)01015-4 docsubty: REV

and Rebelo (1993a, 1993b)], others on financial development [Levine, Loayza and Beck (2000)], and others on macroeconomic policies [Fischer (1993)]. Dollar (1992) stressed a measure of real exchange rate overvaluation as a proxy for outward orientation and thus a determinant of growth. These papers have at least one common feature: they all find that *some* indicator of national policy is strongly linked with economic growth, which confirms the argument made by Levine and Renelt (1992) – even though Levine and Renelt found that it was difficult to discern *which* policy matters for growth. The list of national economic policies that have received most extensive attention are fiscal policy, inflation, black market premiums on foreign exchange, financial repression vs. financial development, real overvaluation of the exchange rate, and openness to trade. The recommendation that countries pursue good policies on all these dimensions was labeled by Williamson (1985) as the "Washington Consensus".

I distinguish policies from "institutions", which have their own rich literature [see Acemoglu, Johnson and Robinson (2001, 2002), La Porta et al. (1999, 1998), Kaufmann, Kraay and Zoido-Lobatón (1999), Levine]. Institutions reflect deep-seated social arrangements like property rights, rule of law, legal traditions, trust between individuals, democratic accountability of governments, and human rights. Although governments can slowly reform institutions, they are not "stroke of the pen" reforms like changes in the macroeconomic policies listed above. I will consider at the end the relative role of policies and institutions in development.

Some empirical caveats

There are several things to note about the evidence on policies and growth before proceeding to new empirical analysis. The first is that the literature has devoted much effort to the most obvious candidate for a policy that influences growth – tax rates. Yet the literature has generally failed to find a link between income or output taxes and economic growth [Easterly and Rebelo (1993a, 1993b), Slemrod (1996)]. Nor are we likely to find that taxes have level effects, as rich countries have higher tax rates than poor countries. The outcome of natural experiments like the large tax increases in the US associated with the introduction of the income tax and the World Wars does not indicate income or level effects of taxes [Rebelo and Stokey (1995)]. Hence, the most obvious policy variable affecting growth is out of the running from the start.

Second, national economic policies are generally measured over the period 1960–2000, which is when data is available. This is also the period in which countries had independent governments making policy, as opposed to colonial regimes (on which we do not have data). Hence, if policies have an effect on the level or growth rate of income, this would have to show up in the period 1960–2000. However, history did not begin with a clean slate in 1960. The correlation of per capita income in 1960 with per capita income in 1999 is 0.87. Most of countries' relative performance is explained by the point they had already reached by 1960. It follows that the role of post-1960 policies

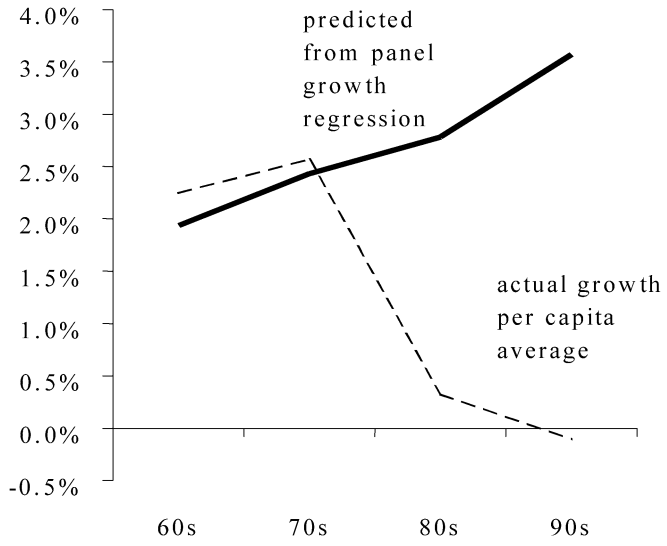


Figure 7. Predicted vs actual per capita growth for developing countries (assuming constant intercept across decades).

in determining development outcomes can only be limited. A view of economic development that puts all the weight on the 1960–2000 period is ahistorical, assuming away the complex histories of civilizations, conquests, and colonies.

Third, there is the general fact that developing countries had higher growth rates in the period 1960–1979 than in the period 1980–2000. Yet most of the "Washington Consensus" policies were adopted only after 1980. In the pre-1980 days, there was much more of an emphasis on state intervention and import-substituting industrialization, as opposed to the free trade, "get the prices right" approach after 1980. This big fact does not augur well for a strong positive effect of "good policies" on growth, although the growth slowdown after 1980 could have other causes. Easterly (2001a) showed the divergence between improving growth predicted by policies and actual growth outcomes across the 60s, 70s, 80s, and 90s (see Figure 7).

Fourth, there are many income differences within nations – between the sexes, between ethnic groups, and between regions – that cannot be explained by national economic policies. Easterly and Levine show that there are four ethnic–geographic clusters of counties with poverty rates above 35 percent in the US: (1) Counties in the West that have large proportions ($>$ 35%) of native Americans; (2) Counties along the Mexican border that have large proportions ($>$ 35%) of Hispanics; (3) Counties adjacent to the lower Mississippi river in Arkansas, Mississippi, and Louisiana and in the "black belt" of Alabama, all of which have large proportions of blacks ($>$ 35%); (4) Virtually all-white counties in the mountains of eastern Kentucky. The county data did not pick up the well-known inner-city form of poverty, mainly among blacks, be-

aghion v.2005/08/19 Prn:23/08/2005; 8:39 F:aghion1015.tex; VTEX/JK p. 21
aid: 1015 pii: S1574-0684(05)01015-4 docsubty: REV

cause counties that include inner cities also include rich suburbs. An inner city zip code in DC, College Heights in Anacostia, has only one-fifth of the income of a rich zip code (20816) in Bethesda MD. This has an ethnic dimension again since College Heights is 96 percent black and the rich zip code in Bethesda is 96 percent white. The purely ethnic differentials in the US are well known. Blacks earn 41 percent less than whites; Native Americans earn 36 percent less; Hispanics earn 31 percent less; Asians earn 16 percent more.[10] There are also more subtle ethnic earnings differentials. Third-generation immigrants with Austrian grandparents had 20 percent higher wages in 1980 than third-generation immigrants with Belgian grandparents [Borjas (1992)]. Among Native Americans, the Iroquois earn almost twice the median household income of the Sioux. Other ethnic differentials appear by religion. Episcopalians earn 31% more income than Methodists [Kosmin and Lachman (1993), p. 260]. Twenty-three percent of the Forbes 400 richest Americans are Jewish, although only two percent of the US population is Jewish [Lipset (1997)].[11]

Poverty areas exist in many countries: northeast Brazil, southern Italy, Chiapas in Mexico, Balochistan in Pakistan, and the Atlantic provinces in Canada. Bouillon, Legovini and Lustig (1999) find that there is a negative Chiapas effect in Mexican household income data, and that this effect has gotten worse over time. Households in the poor region of Tangail/Jamalpur in Bangladesh earned less than identical households in the better off region of Dhaka [Ravallion and Wodon (1998)]. Ravallion and Jalan (1996) and Jalan and Ravallion (1997) likewise found that households in poor counties in southwest China earned less than households with identical human capital and other characteristics in rich Guangdong Province.

In Latin America, the main ethnic divide is between indigenous and non-indigenous populations and between white, mestizo, and black populations. In Mexico, 80.6 percent of the indigenous population is below the poverty line, while only 18 percent of the non-indigenous population is below the poverty line.[12] But even within indigenous groups in Latin America, there are ethnic differentials. There are 4 main language groups among Guatemala's indigenous population. Patrinos (1997) shows that the Quiche-speaking indigenous groups in Guatemala earn 22 percent less on average than Kekchi-speaking groups.

In Africa, there are widespread anecdotes about income differentials between ethnic groups, but little hard data. The one exception is South Africa. South African whites

[10] Tables 52 and 724, 1995 Statistical Abstract of US.

[11] Ethnic differentials are also common in other countries. The ethnic dimension of rich trading elites is well-known: the Lebanese in West Africa, the Indians in East Africa, and the overseas Chinese in Southeast Asia. Virtually every country has its own ethnographic group noted for their success. For example, in the Gambia a tiny indigenous ethnic group called the Serahule is reported to dominate business out of all proportion to their numbers – they are often called "Gambian Jews". In Zaire, Kasaians have been dominant in managerial and technical jobs since the days of colonial rule – they are often called "the Jews of Zaire" (New York Times, 9/18/1996).

[12] Source: [Psacharopoulos and Patrinos (1994, p. 6)].

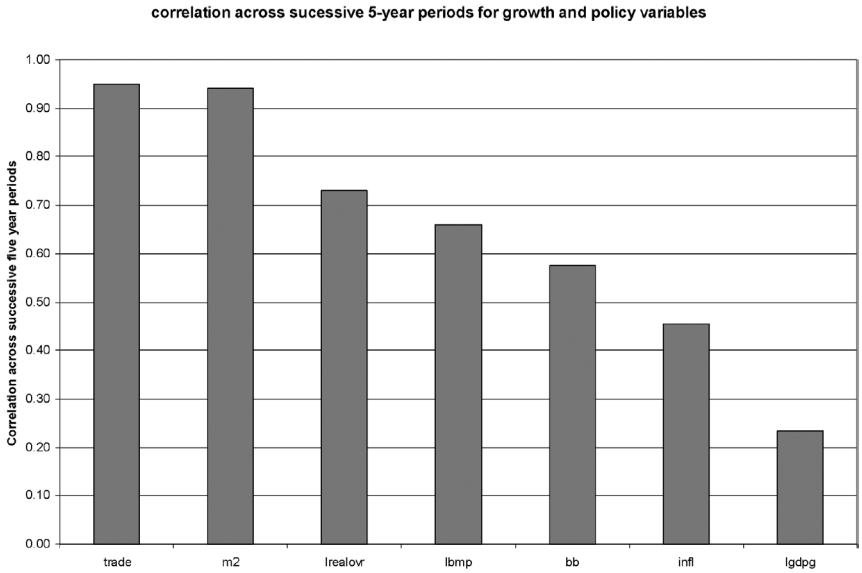


Figure 8. Persistence over time of policies and growth.

have 9.5 times the income of blacks. More surprisingly, among all-black traditional authorities (an administrative unit something like a village) in the state of KwaZulu-Natal, the ratio of the richest traditional authority to the poorest is 54 [Klitgaard and Fitschen (1997)]. While not ruling out national policy effects, these differences also highlight the importance of factors that do not operate at the national level.

Fifth, the role of policies in explaining post-1960 growth is bounded once we realize that policy variables are much more stable over time than are growth rates.[13] Figure 8 shows the correlation coefficient across successive 5-year periods between different kinds of policies and growth. As noted in the theoretical section, stability of policies over time and instability of growth rates is inconsistent with the AK model. It could be consistent with either the neoclassical model or the increasing returns growth model, assuming that policies are close to the steady state or critical point, respectively. Note that the non-persistence of growth rates and the high persistence of income levels is consistent, since persistent differences in growth rates would be required to scramble the income rankings from 1960 to 1999.

New empirical work

I here synthesize past results by running new regressions on an updated dataset for the years 1960–2000, using a panel of five year averages. Following the literature, I con-

[13] This was pointed out by Easterly et al. (1993).

Table 1
Variables used in analysis

Variable name	Definition	Source
LGDPG	Log per capita growth rate	World Bank 2002
INFL	Log (1 + inflation rate)	World Bank 2002
BB	Government budget balance/GDP	World Bank 2002
M2	M2/GDP	World Bank 2002
LREALOVR	Log (overvaluation index/100) (above zero indicates overvaluation)	World Bank 2002
LBMP	Log (1 + black market premium on foreign exchange)	World Bank 2002
TRADE	(Exports+Imports)/GDP	World Bank 2002
GOVC	Government consumption/GDP	World Bank 2002
PRIV	Private sector credit/total credit	World Bank 2002
LNEWGDP	Log of per capita GDP	Summers–Heston 1991 updated using LGDPG
LTYR	Log of total schooling years	Barro–Lee 2000

centrate on the most common measures of macroeconomic policies, price distortions, financial development, and trade openness. My variables are listed in Table 1.

Table 2 shows the variables' summary statistics.

Table 3 shows the correlation coefficients between these variables and growth as well as between distinct policies. All of the bivariate correlations of policy variables with per capita growth are statistically significant at the 5 percent level. Most of the pairwise correlations between policy variables are also statistically significant, indicating the problem of collinearity that has plagued the literature. Bad policies tend to go together along a number of dimensions. M2 and PRIV have such a high correlation that it is clear they are measuring the same thing – the overall level of financial development.

Table 2
Summary statistics

Variable	Obs.	Mean	Standard deviation	Min	Max
INFL	967	0.159	0.325	−0.569	3.447
LNEWGDP	921	8.107	1.040	5.775	10.445
LGDPG	1306	0.017	0.051	−0.736	0.276
GOVC	1241	15.790	6.700	3.915	58.310
BB	958	−0.037	0.054	−0.417	0.391
M2	1064	0.349	0.253	0.009	1.929
PRIV	916	0.355	0.329	0.000	2.085
LREALOVR	609	0.060	0.387	−1.206	1.612
LBMP	1024	0.254	0.558	−1.058	8.311
Trade	1270	0.702	0.454	0.018	3.803
LTYR	832	1.277	0.820	−2.453	2.476

Table 3
Correlation coefficients

	LGDPG	INFL	BB	LREALOVR	LBMP	M2	Trade	PRIV	GOVC
LGDPG	1.000	−0.376	0.155	−0.213	−0.321	0.097	0.101	0.130	−0.130
INFL	−0.376	1.000	−0.201	0.078	0.287	−0.193	−0.078	−0.212	0.031
BB	0.155	−0.201	1.000	−0.141	−0.144	−0.010	0.094	0.110	−0.231
LREALOVR	−0.213	0.078	−0.141	1.000	0.247	−0.083	−0.056	−0.028	0.228
LBMP	−0.321	0.287	−0.144	0.247	1.000	−0.073	−0.178	−0.241	−0.036
M2	0.097	−0.193	−0.010	−0.083	−0.073	1.000	0.375	0.716	0.246
Trade	0.101	−0.078	0.094	−0.056	−0.178	0.375	1.000	0.161	0.276
PRIV	0.130	−0.212	0.110	−0.028	−0.241	0.716	0.161	1.000	0.215
GOVC	−0.130	0.031	−0.231	0.228	−0.036	0.246	0.276	0.215	1.000

I now concentrate on a core set of six variables that seem to capture distinct dimensions of policy: inflation, budget balance, real overvaluation, black market premium, financial depth, and trade openness. Initially, I will test the AK model's prediction that these policies will have growth rather than level effects, so I do not control for initial income (I will check this later on). I will use a variety of specifications and econometric methods to assess how robust are the statistical associations between policies and growth.

I start off with a figure emphasizing the bivariate association between growth and different policies (Figure 9). I divide the sample into two parts, picking out the minority part of the sample where policy is extremely bad and comparing it to the rest (for inflation, black market premium, real overvaluation, and budget balance). Inflation, black market premium, and budget balance all have a distribution featuring a long tail of extreme "bad policy", which seems like a real world experiment worth investigating. So I eyeball the distribution and pick a threshold that picks out this tail of bad policy.

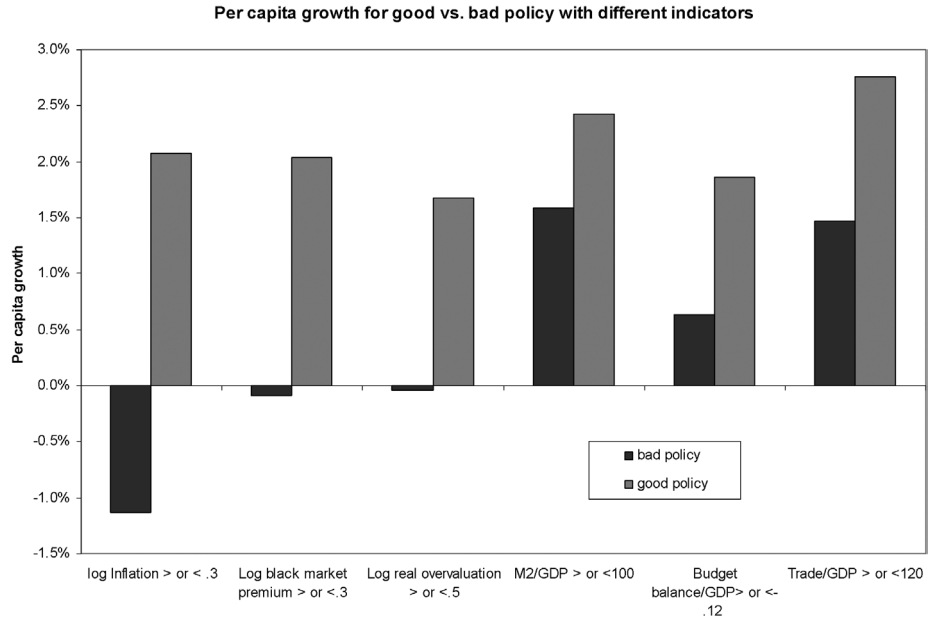


Figure 9. Bivariate effects of policy on growth.

Trade/GDP and M2/GDP have a long tail for extremely *good* policy, so I pick a threshold picking out the extremes of good policy (see Figures 10–15). Real overvaluation does not have a long tail in one direction or the other, but I follow the same practice as with inflation, black market premium, and budget balance in setting a threshold that picks out extremely bad policy. Figure 9 shows that these experiments of either extremely good or extremely bad policy are associated with important growth differences. All of the differences are statistically significant except for the results on M2/GDP. Such strong associations have contributed to the conventional wisdom that policy has strong growth effects.

In Table 4, I regress growth on all six policy variables, and then try dropping one at a time. In the base specification, four of the six policies are statistically significant at the 5 percent level, with trade openness just barely falling short. When I experiment with dropping one variable at a time, all of the six policy variables are significant at one time or another. The coefficients on the policy variables are fairly stable across different permutations of the variables.[14]

Table 5 shows the effect on growth of a one standard deviation improvement in each of the policy variables on growth. If all six variables were improved at the same time,

[14] The other policy variables that I tested: government consumption and private sector credit, were not significant when entered in addition to these variables (or substituting government consumption for budget deficits and private sector credit for M2).

aghion v.2005/08/19 Prn:23/08/2005; 8:39 F:aghion1015.tex; VTEX/JK p. 26
aid: 1015 pii: S1574-0684(05)01015-4 docsubty: REV

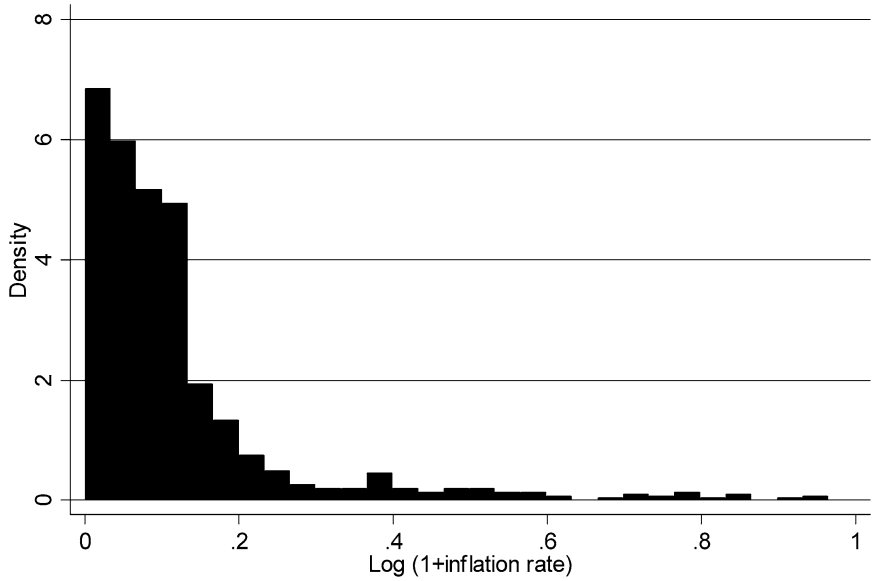


Figure 10. Histogram of inflation (truncated between 0 and 1).

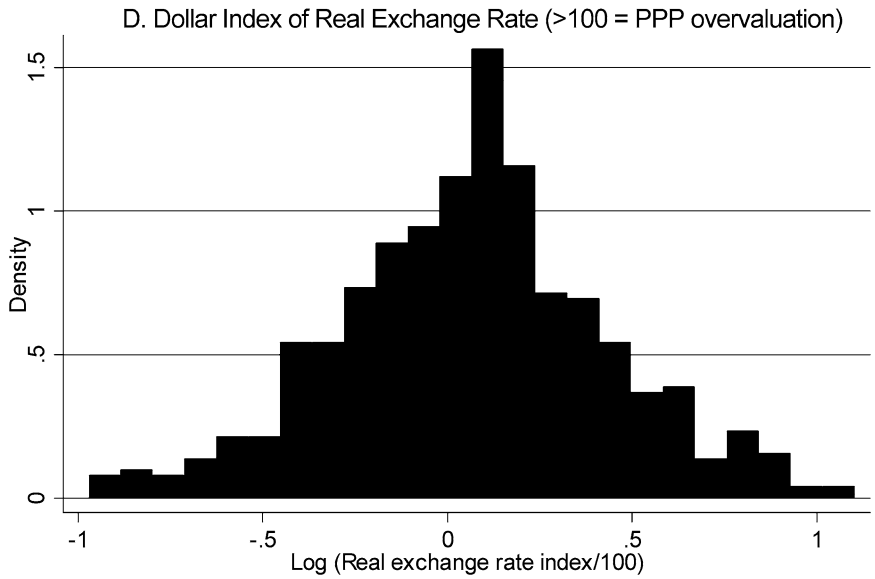


Figure 11. Histogram of real overvaluation (truncated between −1 and 1).

aghion v.2005/08/19 Prn:23/08/2005; 8:39 F:aghion1015.tex; VTEX/JK p. 27
aid: 1015 pii: S1574-0684(05)01015-4 docsubty: REV

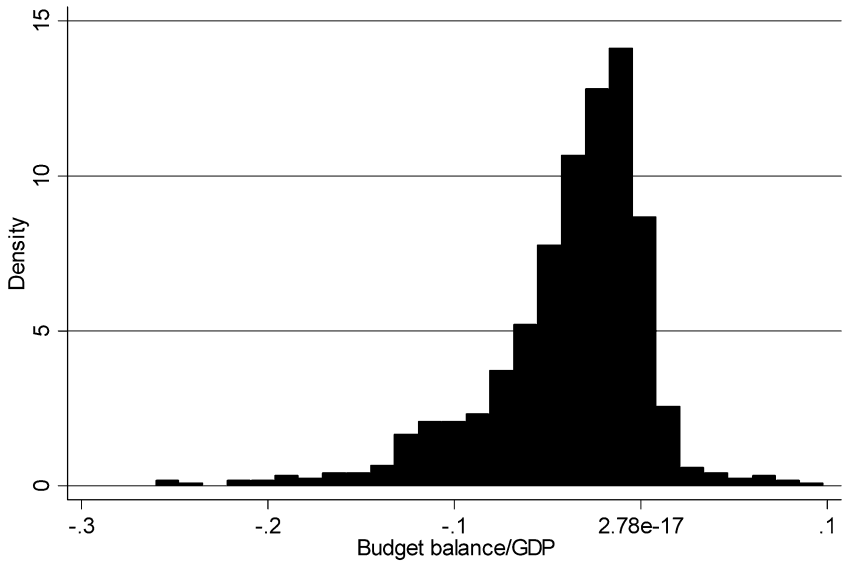


Figure 12. Histogram of budget balance/GDP.

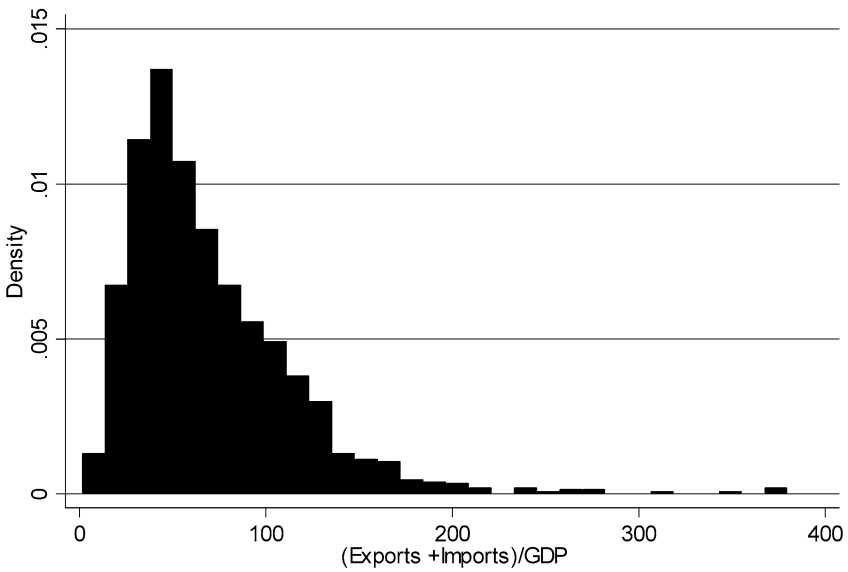


Figure 13. Histogram of trade/GDP (percent).

aghion v.2005/08/19 Prn:23/08/2005; 8:39 F:aghion1015.tex; VTEX/JK p. 28
aid: 1015 pii: S1574-0684(05)01015-4 docsubty: REV

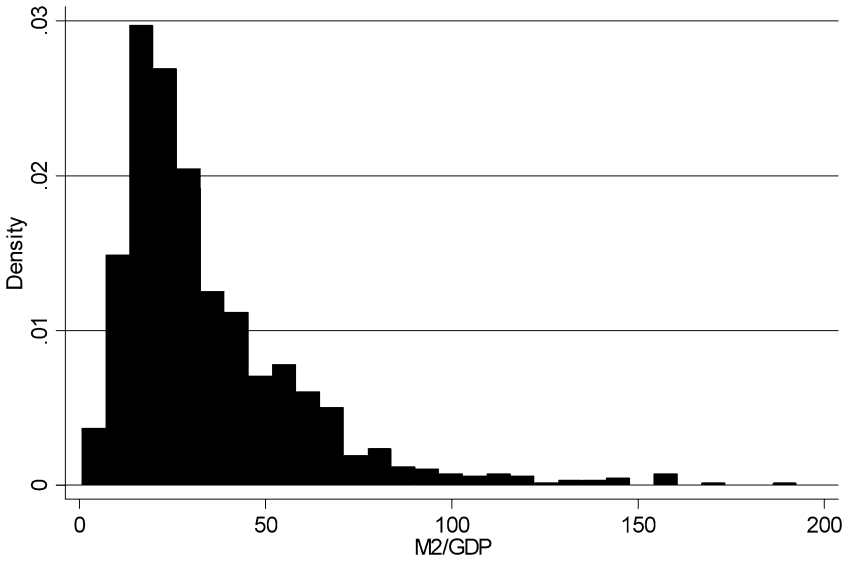


Figure 14. Histogram of M2/GDP (percent).

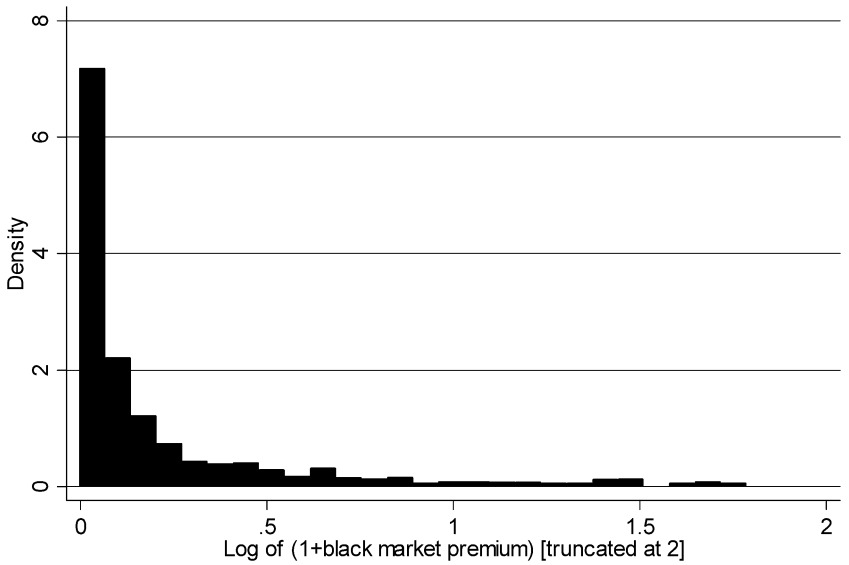


Figure 15. Histogram of Log of (1 + black market premium).

aghion v.2005/08/19 Prn:23/08/2005; 8:39 F:aghion1015.tex; VTEX/JK p. 29
aid: 1015 pii: S1574-0684(05)01015-4 docsubty: REV

Table 4
Regressions of per capita growth on basic set of 6 policy variables. Dependent variable: LGDPG (log per capita growth, five year averages, 1960–2000)

INFL	−0.018		−0.02	−0.02	−0.034	−0.021	−0.018
	(2.61)**		(3.13)**	(2.87)**	(6.27)**	(3.39)**	(2.60)**
BB	0.092	0.114		0.092	0.053	0.109	0.098
	(2.81)**	(3.48)**		(3.07)**	(3.07)**	(3.37)**	(2.92)**
M2	0.01	0.013	0.014		0.017	0.013	0.015
	1.37	1.92	(2.04)*		(2.26)*	(1.99)*	(2.15)*
LREALOVR	−0.014	−0.013	−0.016	−0.013		−0.015	−0.013
	(2.97)**	(2.98)**	(3.74)**	(2.83)**		(3.56)**	(2.88)**
LBMP	−0.012	−0.017	−0.01	−0.014	−0.005		−0.013
	(2.33)*	(3.43)*	(2.06)*	(2.73)**	−0.93		(2.60)**
Trade	0.01	0.011	0.011	0.012	0.001	0.008	
	1.92	(2.22)*	(2.15)*	(2.62)**	0.31	(2.13)*	
Constant	0.016	0.013	0.01	0.021	0.019	0.015	0.021
	(3.62)**	(3.09)**	(2.33)*	(5.67)**	(4.81)**	(3.92)**	(5.55)**
Observations	422	434	458	495	573	455	424
R-squared	0.18	0.15	0.16	0.17	0.13	0.17	0.17

Robust standard errors, significant t statistics in parentheses.
*Significant at 5%.
**Significant at 1%.

Table 5
Effect of one standard deviation improvement in each policy variable on economic growth

Variable	Improvement of one standard deviation in policy variable	Coefficient in growth regression	Change in growth from one standard deviation change in policy (%)
INFL	−0.325	−0.018	0.6
BB	0.054	0.092	0.5
M2	0.253	0.010	0.3
LREALOVR	−0.387	−0.014	0.5
LBMP	−0.558	−0.012	0.7
Trade	0.454	0.010	0.5
Sum			3.0

the regression suggests a 3 percentage point improvement in per capita growth. These results also seem to support the assertion that policies have strong effects on per capita growth.

The promise of getting 3 additional percentage points of growth due to a moderate policy reform package is very seductive. However, there is something disquieting about

aghion v.2005/08/19 Prn:23/08/2005; 8:39 F:aghion1015.tex; VTEX/JK p. 30
aid: 1015 pii: S1574-0684(05)01015-4 docsubty: REV

these results upon further reflection. The one standard deviation change in the policy variables is often very large: reduction of 0.32 in log inflation, 5 percentage point improvement in the budget balance as a ratio to GDP, 25 percentage point increase in M2/GDP, reduction of -0.39 in log real overvaluation, reduction of -0.56 in log black market premium, and increase of 45 percentage points in trade/GDP ratio. Such large changes are outside the experience of most countries with moderate inflation, budget deficits, real overvaluation, black market premiums, etc.

The large standard deviations are related to the long tails I mentioned above. Except for the real overvaluation index, all of the policy variables are highly skewed, with most of the sample concentrated at low values and a few very extreme observations. The outlying observations of inflation, budget deficits, and black market premium are realizations of extreme "bad policies". The outlying observations of trade/GDP and M2/GDP are realizations of extreme "good policies". It is econometric commonsense that extreme observations can be very influential in determining statistical significance of right-hand side variables. How do the above regressions do over more moderate ranges of policy variables?

Table 6 shows the effect of restricting the sample to observations where all six policy variables lie in the range of "moderate" policies. Moderate is defined rather arbitrarily by eye-balling the histograms above to determine where are the cutoffs containing the bulk of the sample (the same cutoffs as in Figure 9 above). Nevertheless, the cutoffs would fit a common-sense description of "extremes": inflation and black market premiums more than 0.3 in log terms (35 percent), real overvaluation more than 0.5 (68 percent), budget deficits greater than 12 percent of GDP, M2 to GDP ratios of more than 100 percent, and trade to GDP ratios of more than 120 percent. The results of excluding any observation where any of the six policy variables are "extreme" is striking: all six policy variables become insignificant, and the F-statistic for their joint effect also falls short of significance. This is not to dismiss the evidence for policy effects on growth (reducing the range of the right-hand side variables would be expected to diminish statistical significance). These extremes are far from irrelevant, as observations in which at least one of the six policies was "extreme" account for more than half the sample. However, these results highlight the dependence of the policy and growth evidence on extreme observations of the policy variables. (The significance of extreme values and the insignificance of moderate ones is also consistent with the prediction of the theoretical model on the nonlinear effects of tax-cum-subsidy policies on economic growth.) There is also the possible endogeneity of these extreme policies, which may reflect general institutional or political chaos. The results suggest that countries not undergoing extreme values of these variables do not have strong reasons to expect growth effects of moderate changes in policies.[15]

[15] The empirical literature on inflation has found that inflation only has a negative effect above some threshold level, although there are disagreements as to where that threshold is [Bruno and Easterly (1998), Barro (1995, 1998), Sarel (1996)].

Table 6
Robustness of results to restricting sample to moderate policy range. Dependent variable is LGDPG

Sample	Full	Moderate policies
INFL	−0.018	−0.064
	(2.61)**	−1.23
BB	0.092	0.018
	(2.81)**	0.22
M2	0.01	−0.004
	1.37	0.27
LREALOVR	−0.014	0.001
	(2.97)**	0.06
Trade	0.01	0.01
	1.92	1.09
LBMP	−0.012	−0.038
	(2.33)*	−0.95
Constant	0.016	0.027
	(3.62)**	(2.52)*
Observations	422	193
R-squared	0.18	0.03

Robust *t* statistics in parentheses.
*Significant at 5%.
**Significant at 1%.
Restrictions under moderate policies: INFL between −0.05 and 0.3, BB between −0.12 and 0.02, M2 < 1.0, LREALOVR between −0.5 and 0.5, Trade < 1.20, LBMP between −0.05 and 0.3.

These results are fairly intuitive if we think of destroying growth as a different process from creating growth. It is a lot easier to cut down a tree than to grow one.[16] Countries that pursue destructive policies like high inflation, high black market premium, chronically high budget deficits and other signs of macroeconomic instability are plausible candidates to miss out on growth. However, it doesn't follow that one can create growth with relative macroeconomic stability. The policies are inherently asymmetric – a leader can sow chaos by printing money and controlling the exchange until he gets a hyperinflation and an absurd black market premium. However, the best he can do in the other direction is zero inflation and zero black market premium. The results on policies and growth may simply reflect the potential for destruction from bad policies, not the potential for fostering long run development through good policy.

The only exception to this story is the trade/GDP variable, whose significance depended on "extremely good" policies. Whatever the source of the result on the extreme,

[16] Easterly (2001a) has a chapter "How governments can destroy growth".

Table 7
Results on initial income and schooling

Dependent variable	LGDPG	LGDPG	LGDPG	LGDPG
INFL	−0.018	−0.019	−0.02	−0.019
	(2.61)**	(2.67)**	(2.65)**	(2.85)**
BB	0.092	0.102	0.124	0.107
	(2.81)**	(2.44)*	(2.65)**	(2.57)*
M2	0.010	0.004	0.002	0.006
	1.37	0.41	0.16	0.67
LREALOVR	−0.014	−0.014	−0.013	−0.014
	(2.97)**	(3.07)**	(2.40)*	(2.96)**
Trade	0.01	−0.01	−0.01	−0.011
	1.92	−1.83	−1.63	−1.96
LBMP	−0.012	0.01	0.008	0.009
	(2.33)*	1.87	1.37	1.62
LNEWGDP		0.003	−0.001	0.0480
		1.4	−0.28	1.96
LTYR			0.007	
			1.42	
LNEWGDP^2				−0.0030
				−1.87
Constant	0.016	−0.004	0.019	−0.187
	(3.62)**	−0.25	−0.87	−1.86
Observations	422	411	359	411
R-squared	0.18	0.18	0.19	0.18

Robust t statistics in parentheses.
*Significant at 5%.
**Significant at 1%.
Turning point for convergence is 2981.

this suggests that opening up for most economies – who likely would not reach this extreme even under complete free trade – would not be associated with growth effects.

The next thing to test is whether initial income belongs in the growth equation, as the neoclassical model would imply. It has also been common in the literature to add initial schooling as an indicator of whether the balance between physical and human capital is far from the optimal level. Table 7 shows the results on initial income and schooling.

The results are not very supportive of a conditional convergence result. Initial income and schooling do not enter significantly, although a nonlinear formulation of hump-shaped conditional convergence (including initial income squared) comes close to significance.[17] Since there is a large literature starting with Barro (1991) and Barro and Sala-i-Martin (1992) that does find conditional convergence, I do not claim this result

[17] Hump-shaped convergence is consistent with a neoclassical model in which there is some subsistence floor to consumption (the Stone–Geary utility function).

Table 8
Panel methods in policies and growth regressions. Dependent variable: LGDPG

Panel method	Random effects	Between	Fixed effects
INFL	−0.019	−0.012	−0.02
	(3.53)**	−0.97	(3.43)**
BB	0.082	0.216	0.069
	(2.35)*	(3.51)**	−1.64
M2	0.002	0.026	−0.057
	−0.22	(2.19)*	(3.16)**
LREALOVR	−0.009	−0.027	0.01
	−1.8	(3.82)**	−1.43
Trade	0.012	0	0.046
	−1.95	−0.07	(3.19)**
LBMP	−0.011	−0.01	−0.012
	(2.15)*	−0.97	−1.84
Constant	0.017	0.019	0.016
	(3.22)**	(2.73)**	−1.61
Observations	422	422	422
Number of country	88	88	88
R-squared	0.17	0.41	0.13
Sample	Full	Full	Full
Reject random effects	Yes		

Absolute value of z statistics in parentheses.
*Significant at 5%.
**Significant at 1%.

is decisive. It does show the fragility of the results on both policies and initial conditions (note that three of the policy variables become insignificant when initial income is included). I will come back to the issue of conditional convergence when I examine effects of policy on growth with dynamic panel methods.

There is another robustness check that we should perform on the policies and growth results. Following common practice in the literature, I have been doing regressions on pooled time series cross-section observations. This implicitly assumes that the effects on growth of a policy change over time are the same as a policy difference between countries. It is straightforward to test this restriction by doing within and between regressions on the pooled sample. Table 8 shows the results. I also show the results of a random effects regression, which gives results similar to OLS on the pooled sample. The test of whether the random effects are orthogonal to the right-hand side variables is an indirect test of the equality of the coefficients from the between and within regressions. I strongly reject the hypothesis that the random effects are orthogonal. We can see from the between and within (fixed effects) regressions that the coefficients across time and across countries are indeed very different. Inflation is not significant in the be-

aghion v.2005/08/19 Prn:23/08/2005; 8:39 F:aghion1015.tex; VTEX/JK p. 34
aid: 1015 pii: S1574-0684(05)01015-4 docsubty: REV

tween regression but strongly significant in the within regression.[18] The budget balance is the reverse: strongly significant in the between regression but not in the fixed effects regression. The weak result that I found on M2/GDP in the pooled regression turns out to be because the between and within effects tend to cancel out: M2/GDP is strongly positively correlated with growth in the between regression and negatively correlated with growth in the within regressions. Real overvaluation and trade also show different results in the two different panel methods (real overvaluation is significant between countries and insignificant within countries, while trade is the reverse). This instability of growth effects is inconsistent with a simple AK view of growth with instantaneous transitional dynamics. It is also possible that five year averages are not long enough to wipe out cyclical fluctuations. The negative correlation between M2/GDP and growth could be seen as a cyclical pattern such as a loosening of monetary policy during recessions and tightening during booms. Likewise the correlation of trade/GDP with growth could indicate that international trade is pro-cyclical, as opposed to indicating any causal effect of openness on growth.

Also note that the r-squared of the between regression is much higher than the within regression. This is not surprising given that the between regression is on averages, but it does show that the growth effects of most concern to policy makers – the change over time within a given country of growth in response to policy changes – are very imprecisely estimated by the data. Fully 87 percent of the within country variance in growth rates is not explained by these six policy variables. This result is not surprising when we recall the persistence of policies over time and the non-persistence of growth rates.

Another panel method I apply to the data is the well-known dynamic panel estimator of Arellano and Bond (1991). This estimator uses first differences to remove the fixed effects. This method has several advantages: (1) it addresses reverse causality concerns by using twice-lagged values of the right-hand side variables as instruments for the first differences of RHS variables, (2) we can include initial income again, which is not possible with traditional panel methods because it would be correlated with the error term (Arellano and Bond address this by instrumenting for initial income with the twice-lagged value), and (3) we can also include the lagged growth rate to allow for partial adjustment of growth to policy changes, which is more plausible than instantaneous adjustment.

The results in Table 9 are notable in reinvigorating the conditional convergence hypothesis. This is consistent with previous work that shows a higher coefficient (in absolute value) on initial income with dynamic panel methods than with pooled or cross-section OLS [Caselli, Esquivel and Lefort (1995)]. The coefficient on lagged growth is not significant, failing to find support for the partial adjustment hypothesis. The results on policies are similar (not surprisingly) to the fixed effects estimator above. Inflation

[18] This is consistent with the Bruno and Easterly (1998) result that high inflation crises have a strong temporary negative effect on output but no permanent effects.

Table 9
Regressions using Arellano and Bond dynamic panel method

Dependent variable: LGDPG	(1)	(2)	(3)	(4)
LD.LGDPG	−0.0441	−0.1131	−0.09	−0.0627
	(0.0749)	(0.0674)*	(0.0771)	(0.0823)
D.INFL	−0.0137	−0.0141	−0.0162	−0.017
	(0.0068)**	(0.0064)**	(0.0065)**	(0.0066)***
D.BB	0.1014	0.0958	0.0876	0.0544
	(0.0509)**	(0.0501)*	(0.0540)	(0.0571)
D.M2	−0.0701	−0.0457	−0.0522	−0.0486
	(0.0286)**	(0.0284)	(0.0302)*	(0.0307)
D.LREALOVR	0.0085	0.0081	0.0083	0.0021
	(0.0093)	(0.0087)	(0.0092)	(0.0098)
D.LBMP	−0.0084	−0.008	−0.0037	0.0013
	(0.0086)	(0.0081)	(0.0086)	(0.0090)
D.Trade	0.0715	0.072	0.0635	0.0555
	(0.0201)***	(0.0193)***	(0.0204)***	(0.0211)***
D.LNEWGDP		−0.0487	−0.0508	−0.0466
		(0.0098)***	(0.0104)***	(0.0105)***
D.LTYR			0.0091	0.0137
			(0.0104)	(0.0105)
Constant	−0.0012	0.0014	0.0019	0.001
	(0.0016)	(0.0016)	(0.0020)	(0.0021)
Observations	323	316	275	275
Number of country	82	79	69	69
Sargan	36.51018	35.57537	31.19113	23.98194
Prob > CHI2	0.0091	0.0119	0.0385	0.1968
Test first order autocovariance	0	0	0	0
Test second order autocovariance	0.9091	0.4666	0.5212	0.4797
Time dummies	No	No	No	Yes

Standard errors in parentheses.
*Significant at 10%.
**Significant at 5%.
***Significant at 1%.

and trade are strongly significant with the right sign, while M2/GDP still has a significant but perverse sign. The results do not change much if I experiment with omitting one policy variable at a time. The estimates are consistent because I fail to reject that second order serial correlation is zero. The difference with the fixed effects result on policies is that these results have somewhat more claim to being causal. However, the Sargan test rejects the overidentifying restrictions, except in the last equation where I add time dummies. This highlights a weakness of the strong claims for causality made by the dynamic panel method – they depend on the rather dubious assumption that the lagged right-hand side variables do not themselves enter the growth equation. The same problem afflicts the cross-section or pooled regressions that use lagged values of policy

as instruments for current policies. Traditionalists who like intuitive arguments why instruments plausibly affect the independent but not the dependent variable are not very persuaded by lagged policies as instruments. As Mankiw (1995) noted sarcastically, if I instrument for the price of apples with the lagged price of apples in an equation for the quantity of apples, is it the supply or demand equation that I have identified?

Policy episodes and transition paths

A more informal approach to detecting the nature of policy effects on growth is to do episodic analysis – try to identify major policy reforms and simply examine what happened to growth and investment before and after. The shortcomings of this approach are that we do not control for other factors that affect growth and that it is somewhat arbitrary to define what are "major policy reforms". The advantage is that we can see the annual path of growth rates and thus get a better test of the different prediction for post-reform transitional dynamics made by the models in the theoretical section.

One ambitious attempt to identify major policy reform episodes was made by Sachs and Warner (1995). Sachs and Warner rate an economy as closed if any of the following hold: (1) a black market premium more than 20 percent, (2) the government has a purchasing monopoly at below-market prices on a major commodity export, (3) the country has a socialist economic system, (4) non-tariff barriers cover more than 40 percent of intermediate and capital goods imports, and (5) weighted average tariff of more than 40 percent on intermediate and capital goods. Note that only some of these criteria have anything to do with "trade openness" in the usual sense, as pointed out by Rodriguez and Rodrik (2001). The important thing for my purposes is that Sachs and Warner identify the dates of "reform" according to these criteria. I utilize an updated series of Sachs–Warner openness that goes through 1998.[19] I pick out countries with at least 13 years of growth data after opening. Since most openings happen towards the end of the sample period, this limits the sample of countries to only 13: Botswana, Chile, Colombia, Costa Rica, Ghana, Guinea, the Gambia, Guinea-Bissau, Israel, Mexico, Morocco, New Zealand, and Papua New Guinea. Figure 16 shows the path of growth and investment before and after opening, after first smoothing each country's series individually with an HP filter. The results do not support any of the above policies and growth models very convincingly. Investment is completely at variance with the predictions for its transitional path. Growth does show a steady acceleration after opening. This could be either a symptom of increasing returns or simply a process of increased credibility as the reforms take hold. Note however that growth was highest many years before the opening. Perhaps the story of closed and open economies is something more complex like temporary high growth under import substitution, which eventually crashed, followed by an opening of the economy and a partial recovery of growth.

[19] The source is Easterly, Levine and Roodman (2003).

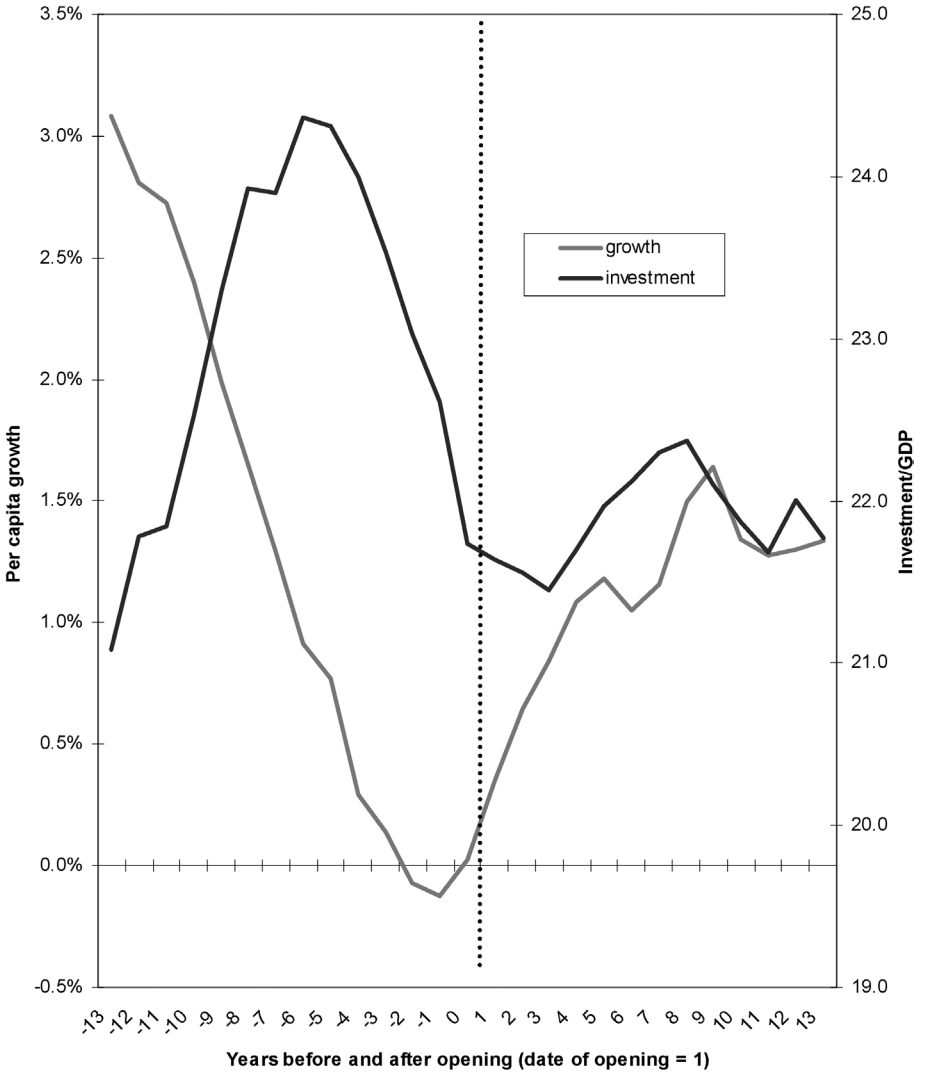


Figure 16. Growth and investment before and after opening economy in 13 countries.

One of the few cases to fit the predictions of growth models as to transitional dynamics is Ghana, where both investment and growth increase after opening. Both keep rising after the date of opening, again supporting either an increasing returns story or increasing credibility of reform (see Figure 17).

Another type of reform that lends itself to transition analysis is stabilization from high inflation. I record episodes of high inflation as following the above definition (log

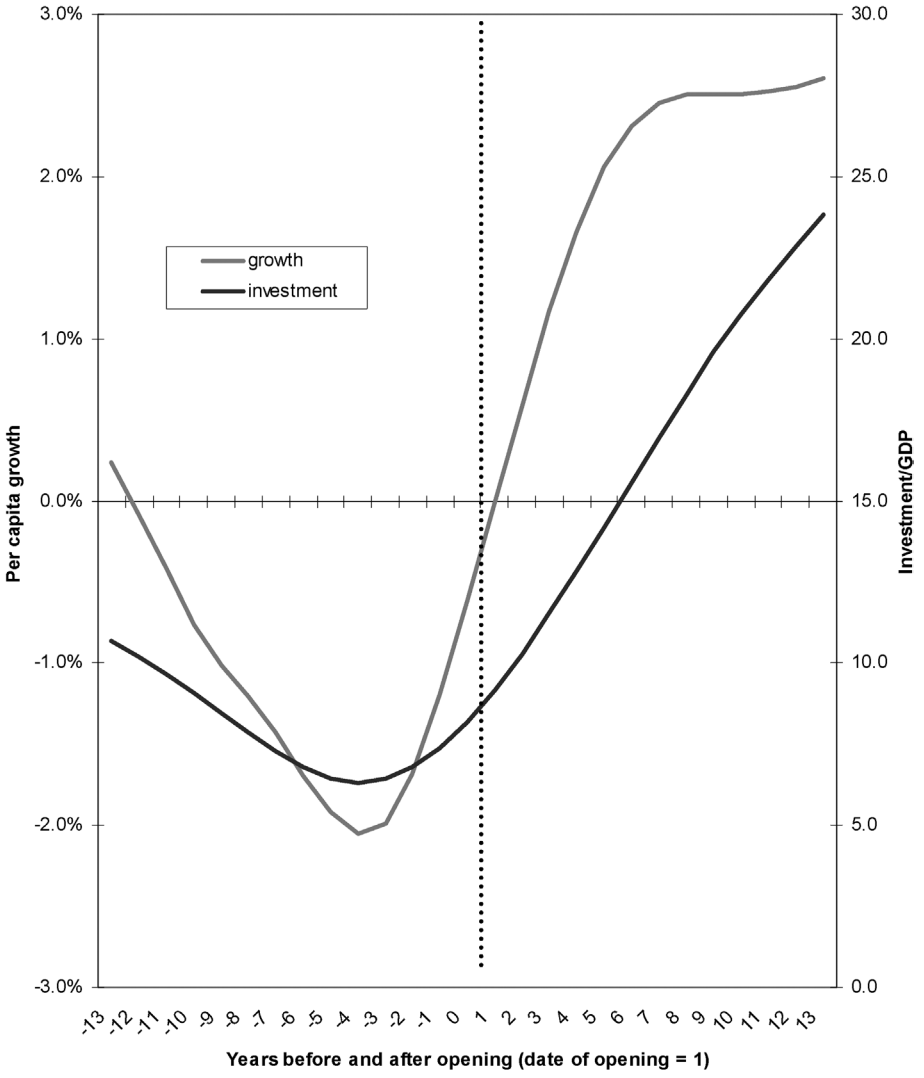


Figure 17. Growth and investment before and after opening economy in Ghana.

rate of inflation above 0.3). I measure years of high inflation prior to stabilization, and then years after stabilization when inflation remains below 0.3. I require that there be at least two years of high inflation to rule out one-time spikes in the price level. The first year after inflation comes down is recorded as year 1. Figure 18 shows the behavior of growth and investment before and after inflation comes down. Growth fits the prediction of theoretical models in jumping to a higher path immediately after inflation comes

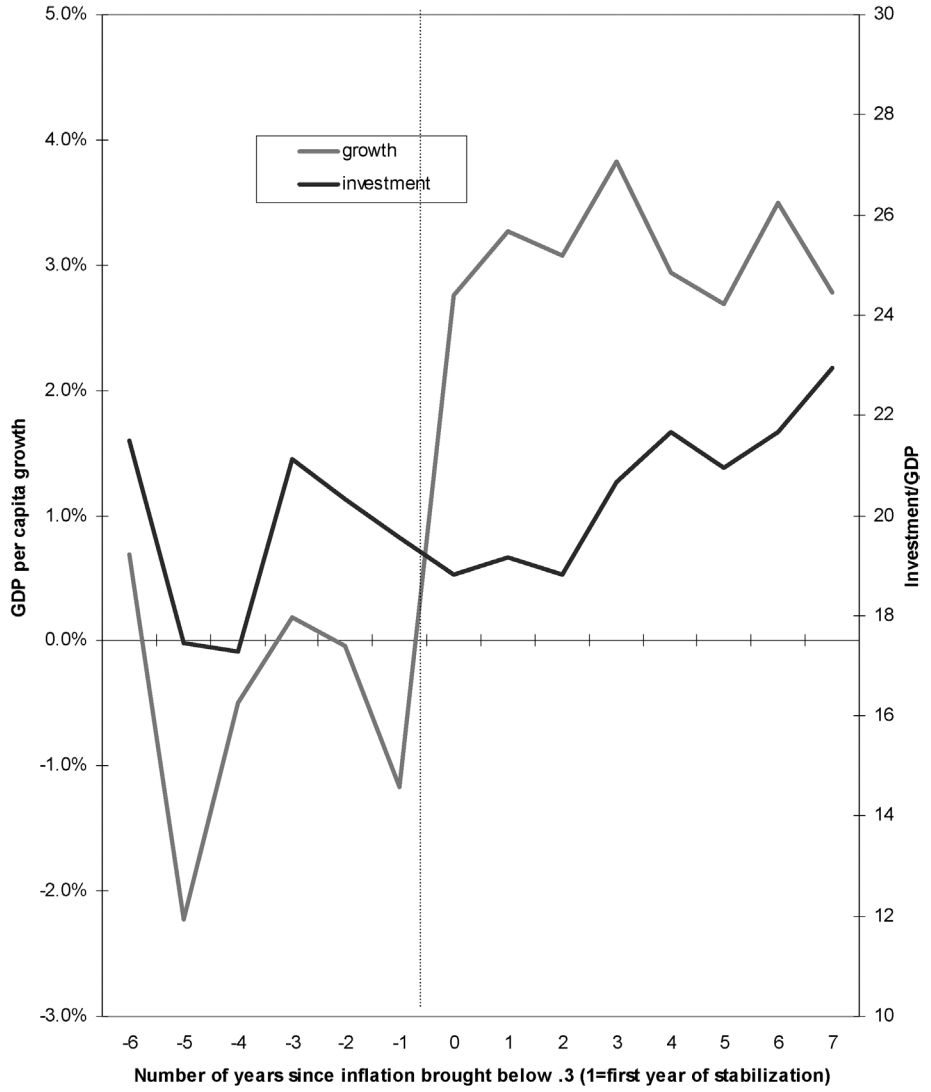


Figure 18. Investment and growth after inflation stabilization.

down. We only have a large enough sample for 7 years after inflation comes down, but growth seems to remain fairly constant post-stabilization. Investment fails to fit the transition predictions of any of the models.

We are left with a somewhat mixed picture. There is a fairly rapid growth effect after policy reform, either accelerating or constant. Investment in physical capital does not seem to respond to reforms in the way predicted by growth models. Of course,

causality is up for grabs. There is also still the extreme policies problem, as episodes in which the country was closed or inflation was very high reflect asymmetrically destructive policies; it is not surprising that growth rebounds after these policies are terminated.

Institutions versus policies

Recent research has examined the relative role of historical institutions and more recent government policy behavior. The *institutions view* holds that geographic and historical conditions produce long-lasting differences in institutions. For example, environments where crops are most effectively produced using large plantations will quickly develop political and legal institutions that protect the few landholders from the many peasants [Engermann and Sokoloff (1997)]. Even when agriculture recedes from the economic spotlight, enduring institutions will continue to thwart competition and hence economic development. Similarly, many countries' institutions were shaped during colonization, so that examining colonies is a natural experiment. European colonialists found different disease environments around the globe. In colonies with inhospitable germs and climates, the colonial powers established extractive institutions, so that a few colonialists could exploit natural resources. In colonies with hospitable climates and germs, colonial powers established settler institutions. According to this view, the institutional structures created by the colonialists in response to the environment endure even with the end of colonialism [Acemoglu, Johnson and Robinson (2001, 2002)]. A history of ethnolinguistic divisions may both prevent the development of good institutions and be more damaging when those institutions are absent [Mauro (1995), Easterly and Levine (1997), and Easterly (2001a)]. Thus, the institution view argues that economic development mainly depends on institutions that reflect deep-seated historical factors [North (1992)].

In contrast, the *policy* view – which is really a collection of many different approaches – questions the importance of history or geography in shaping economic development today. This view is embedded in the approach of multilateral development institutions. The policy view holds that economic policies and institutions reflect current knowledge and political forces. Thus, changes in either knowledge about which policies and institutions are best for development or changes in political incentives will produce rapid changes in institutions and economic policies. According to the policy view, while history and geography may have influenced production and institutions, understanding them is not crucial to understanding economic development today.

Easterly and Levine (2003) examine whether major macroeconomic policies – inflation, trade policies, and impediments to international transactions as reflected in real exchange rate overvaluation – help explain current levels of economic development, after controlling for institutions. They do this in two steps. First, they treat the macroeconomic policy indicators, which are averaged over the last four decades as exogenous.

Simultaneity bias may bias these results toward finding a significant statistical relationship between policies and economic development if economic success tends to produce better policies. Second, they treat the macroeconomic policy indicators as endogenous; they use instrumental variables (geographic variables and ethnolinguistic fractionalization) to control for potential simultaneity bias. Using these two methods, they assess whether macroeconomic policies explain cross-country differences in economic development. In both methods they instrument for institutions with the set of variables discussed above.

The evidence suggests that macroeconomic policies do not have a significant impact on economic development after accounting for the impact of institutions on the level of economic development. When the policy variables are treated as included exogenous variables, the Institutions Index enters all of the regression significantly. Furthermore, the coefficient size on the Institutions Index is essentially unchanged from regressions that did not include policy indicators. Thus, even after controlling for macroeconomic policies, institutions explain cross-country differences in economic development. Furthermore, the data never reject the OIR-test. The policy indicators never enter the regressions significantly. Inflation, Openness, and Real Exchange Rate Overvaluation never enter with a P-value below 0.10. Moreover, even when they are included together, the data do not reject the null hypothesis that the three policies all enter with coefficients equal to zero, which is shown using the F-test on the three policy variables.

When using instrumental variables for the policy indicators, they again find that macroeconomic policies do not explain economic development. Specifically, they fail to reject that hypothesis that macroeconomic policies have zero impact on economic development after accounting for the impact of institutions.

As noted earlier, the instrumental variables explain a significant amount of the cross-country variation in the Institutions Index. In the first-stage regressions for policy, Easterly and Levine (2003) find that the instruments explain a significant amount of the cross-country variation in Openness and Real Exchange Rate Overvaluation at the 0.01 significance level. However, the instruments do not do a very good job of explaining cross-country variation in inflation, i.e., they fail to find evidence that the instruments explain average inflation rates over the last four decades at the 0.01 significance level. The policy variables never enter significantly in either method. While the exogenous component of the Institutions Index (i.e., the component defined by endowments) continues to significantly account for international differences in the level of GDP per capita, the macroeconomic policy indicators do not add any additional explanatory power.

This raises the suspicion that adverse macroeconomic policies (and macroeconomic volatility in general) may have been proxying for poor institutions in growth regressions. Acemoglu et al. (2003) provide some evidence supporting this suspicion.

In sum, the long run effect of policies on development is difficult to discern once you also control for institutions.

Conclusions

The large literature on national policies and growth established some statistical association between national economic policies and growth. I confirm that association in this paper and I show how it could have reasonable theoretical foundations. However, I find that the associations seem to depend on extreme values of the policy variables, that the results are not very robust to different econometric methods or introducing initial income, and that a levels regression does not show any effect of policies after controlling for institutions (both instrumented for possible endogeneity). These results are consistent with other theoretical models that predict only modest effects of national policies, depending on model parameters, and show nonlinear effects of tax-cum-subsidy schemes. They are also consistent with the view that the residual A explains most of income and growth differences, and it likely reflects deep-seated institutions that are not very amenable to change in the short run.

Uncited references

[Barro and Sala-i-Martin (1995)] [Barro et al. (1995)] [Beck, Levine and Loayza (2000)] [Blomstrom, Lipsey and Zejan (1996)] [Boyd, Ross and Smith (2001)] [Burnside (1996)] [Carroll and Weil (1993)] [Chari, Christiano and McGrattan (1996)] [De Gregorio (1992)] [De Gregorio (1993)] [Durlauf and Quah (in preparation)] [Easterly (1999)] [Easterly (2001b)] [Hsieh (1998)] [Jones (1995a)] [Jones (1995b)] [Jones (1997)] [King and Levine (1994)] [King and Rebelo (1993)] [Krugman and Venables (1995)] [Levine and Zervos (1993)] [Levine and Zervos (1998)] [Lucas (1988)] [Lucas (1990)] [Lucas (1998)] [Maddison (1995)] [McGrattan (1998)] [Parente (1994)] [Parente and Prescott (1996)] [Prescott (1998)] [Pritchett (1998)] [Pritchett (1999)] [Psacharopoulos (1994)] [Quah (1993)] [Rauch (1993)] [Ray (1998)] [Rebelo (1998)] [Rodrik (1998)] [Romer (1990)] [Shleifer and Vishny (1993)] [Sokoloff and Engerman (2000)] [Solow (1956)] [Solow (1957)]

References

Acemoglu, D., Johnson, S., Robinson, J. (2001). “The colonial origins of comparative development”. American Economic Review.

Acemoglu, D., Johnson, S., Robinson, J. (2002). “Reversal of fortunes: Geography and institutions in the making of the modern world income distribution”. Quarterly Journal of Economics 117 (in press).

Acemoglu, D., Johnson, S., Robinson, J., Thaicharoen, Y. (2003). Institutional Causes, Macroeconomic Symptoms: Volatility, Crises and Growth. Journal of Monetary Economics (in press).

Ades, A., Glaeser, E. (1999). “Evidence on growth, increasing returns, and the extent of the market”. Quarterly Journal of Economics CXIV (3), 1025–1046.

Aghion, P., Howitt, P. (1998). Endogenous Growth Theory. MIT Press.

Arellano, M., Bond, S. (1991). “Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations”. Review of Economic Studies 58 (2), 277–297.

Azariadis, C., Drazen, A. (1990). "Threshold externalities in economic development". Quarterly Journal of Economics 105, 501–526.

Barro, R.J. (1991). "Economic growth in a cross section of countries". Quarterly Journal of Economics 106 (2), 407–443.

Barro, R.J. (1995). "Inflation and economic growth". Bank of England Quarterly Bulletin, May, 166–176.

Barro, R.J. (1998). "Determinants of Economic Growth: A Cross Country Empirical Study". MIT Press, Cambridge, MA.

Barro, R., Sala-i-Martin, X. (1992). "Convergence". Journal of Political Economy.

Barro, R., Sala-i-Martin, X. (1995). Economic Growth. McGraw-Hill, New York.

Barro, R., Mankiw, J., Gregory, N., Sala-i-Martin, X. (1995). "Capital mobility in neoclassical models of growth". American Economic Review 85 (March), 103–115.

Beck, T., Levine, R., Loayza, N. (2000). "Finance and the sources of growth". Journal of Financial Economics 58 (1–2), 261–300.

Becker, G.S., Murphy, K.M., Tamura, R. (1990). "Human capital, fertility, and economic growth". Journal of Political Economy 98 (5), S12–S37. Part 2.

Benabou, R. (1993). "Workings of a city: Location, education, and production". Quarterly Journal of Economics 108 (August), 619–652.

Bénabou, R. (1996). "Heterogeneity, stratification, and growth: Macroeconomic implications of community structure and school finance". American Economic Review 86 (3), 584–609.

Benhabib, J., Spiegel, M. (1994). "Role of human capital in economic development: Evidence from aggregate cross-country data". Journal of Monetary Economics 34, 143–173.

Blomstrom, M., Lipsey, R., Zejan, M. (1996). "Is fixed investment the key to economic growth?". Quarterly Journal of Economics, February, 269–276.

Borjas, G.J. (1992). "Ethnic capital and intergenerational mobility". Quarterly Journal of Economics 107 (February), 123–150.

Borjas, G.J. (1995). "Ethnicity, neighborhoods, and human capital externalities". American Economic Review 85 (3), 365–390.

Borjas, G.J. (1999). Heaven's Door: Immigration Policy and the American Economy. Princeton University Press, Princeton.

Boyd, J.H., Ross, L., Smith, B.D. (2001). "The impact of inflation on financial sector performance". Journal of Monetary Economics (in press).

Bruno, M., Easterly, W. (1998). "Inflation crises and long-run growth". Journal of Monetary Economics 41, 3–26.

Burnside, C. (1996). "Production function regressions, returns to scale and externalities". Journal of Monetary Economics 37, 177–200.

Carroll, C.D., Weil, D.N. (1993). "Saving and Growth: A Reinterpretation". Carnegie–Rochester Series on Public Policy.

Chari, V.V., Christiano, L., McGrattan, E. (1996). "The Poverty of Nations: A Quantitative Exploration". NBER Working Paper No. 5414.

Collier, P., Dollar, D. (2001). "Can the world cut poverty in half? How policy reform and international aid can meet international development goals". World Development.

De Gregorio, J. (1992). "The effects of inflation on economic growth". European Economic Review 36 (2–3), 417–424.

De Gregorio, J. (1993). "Inflation, taxation and long-run growth". Journal of Monetary Economics 31, 271–298.

Dollar, D. (1992). "Outward-oriented developing economies really do grow more rapidly: Evidence from 95 LDCs, 1976–1985". Economic Development and Cultural Change 40 (3), 523–544.

Durlauf, S., Quah, D. "The new empirics of economic growth." In: Taylor, J., Woodford, M. (Eds), Handbook of Macroeconomics (in preparation).

Easterly, W. (1993). "How much do distortions affect growth?". Journal of Monetary Economics 32 (November), 187–212.

Easterly, W. (1994). "Economic stagnation, fixed factors, and policy thresholds". Journal of Monetary Economics 33, 525–557.

Easterly, W. (1999). "The ghost of financing gap: Evaluating the growth model of the international financial institutions". Journal of Development Economics, December.

Easterly, W. (2001a). "The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics". MIT Press, Cambridge MA. Paperback edition 2002.

Easterly, W. (2001b). "The lost decades: Developing countries' stagnation in spite of policy reform 1980–98". Journal of Economic Growth.

Easterly, W., Levine, R. (1997). "Africa's growth tragedy: Policies and ethnic divisions". Quarterly Journal of Economics, November.

Easterly, W., Levine, R. (2001). "It's not factor accumulation: Stylized facts and growth models". World Bank Economic Review.

Easterly, W., Levine, R. (2003). "Tropics, germs, and crops: The role of endowments in economic development". Journal of Monetary Economics 50 (1).

Easterly, W., Rebelo, S. (1993a). "Fiscal policy and economic growth: An empirical investigation". Journal of Monetary Economics 32, 417–458.

Easterly, W., Rebelo, S. (1993b). "Marginal income tax rates and economic growth in developing countries". European Economic Review 37, 409–417.

Easterly, W., Kremer, M., Pritchett, L., Summers, L. (1993). "Good policy or good luck? Country growth performance and temporary shocks". Journal of Monetary Economics 32 (December), 459–483.

Engermann, S., Sokoloff, K. (1997). "Factor endowments, institutions, and differential paths of growth among new world economies: A view from economic historians of the United States". In: Haber, S. (Ed.), How Latin America Fell behind. Stanford University Press, Stanford, CA.

Fischer, S. (1993). "The role of macroeconomic factors in growth". Journal of Monetary Economics 32, 485–512.

Frankel, J.A., Romer, D. (1999). "Does trade cause growth?". American Economic Review 89, 379–399.

Fujita, M., Krugman, P., Venables, A. (1999). The Spatial Economy: Cities, Regions, and International Trade.

Hall, R.E., Jones, C. (1999). "Why do some countries produce so much more output per worker than others?". Quarterly Journal of Economics 114 (1), 83–116.

Hsieh, C.-T. (1998). "What explains the industrial revolution in East Asia? Evidence from factor markets". Princeton University Woodrow Wilson School of Public and International Affairs Discussion Papers in Economics No. 196 (January) 1–42.

International Monetary Fund (2000). Policies for Faster Growth and Poverty Reduction in Sub-Saharan Africa and the Role of the IMF. Issues Brief. Washington, DC.

Jalan, J., Ravallion, M. (1997). "Spatial poverty traps?". World Bank Policy Research Working Paper No. 1862.

Jones, C. (1995a). "R&D-based models of economic growth". Journal of Political Economy 103 (August), 759–784.

Jones, C. (1995b). "Time series tests of endogenous growth models". Quarterly Journal of Economics 105 (2), 495–526.

Jones, C. (1997). "Comment on Peter Klenow and Andres Rodriguez-Clare, "The neoclassical revival in growth economics: Has it gone too far?" ". NBER Macroeconomics Annual 12, 73–103.

Kaufmann, D., Kraay, A., Zoido-Lobatón, P. (1999). "Governance Matters". World Bank Research Working Paper 2196.

King, R.G., Levine, R. (1994). "Capital fundamentalism, economic development and economic growth". Carnegie–Rochester Conference Series on Public Policy 40, 259–292.

King, R.G., Rebelo, S. (1993). "Transitional dynamics and economic growth in the neoclassical model". American Economic Review 83, 908–931.

Klenow, P. (1998). "Ideas versus rival human capital: Industry evidence on growth models". Journal of Monetary Economics 42, 2–23.

Klenow, P., Rodriguez-Clare, A. (1997a). "Economic growth: A review essay". Journal of Monetary Economics 40, 597–617.

Klenow, P., Rodriguez-Clare, A. (1997b). "The neoclassical revival in growth economics: Has it gone too far?". NBER Macroeconomics Annual 1997 12, 73–103.
Klitgaard, R., Fitschen, A. (1997). "Exploring income variations across traditional authorities in KwaZulu-Natal, South Africa". Development Southern Africa 14 (3).
Kosmin, B.A., Lachman, S.P. (1993). "One Nation under God: Religion in Contemporary American Society". Harmony Books.
Kremer, M. (1993). "O-ring theory of economic development". Quarterly Journal of Economics 108 (August), 551–575.
Krugman, P.R. (1991). Geography and Trade. MIT Press, Cambridge, MA.
Krugman, P. (1995). Development, Geography, and Economic Theory. MIT Press, Cambridge, MA.
Krugman, P. (1998). "Space: The final frontier". Journal of Economic Perspectives 12 (Spring), 161–174.
Krugman, P.R., Venables, A.J. (1995). "Globalization and the inequality of nations". Quarterly Journal of Economics, November, 857–880.
La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R.W. (1998). "Law and finance". Journal of Political Economy 106, 1113–1155.
La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R.W. (1999). "The quality of government". Journal of Law, Economics, and Organization 15, 222–279.
Levine, R., Loayza, N., Beck, T. (2000). "Financial intermediation and growth: Causality and causes". Journal of Monetary Economics 46 (August), 31–77.
Levine, R., Renelt, D. (1992). "A sensitivity analysis of cross-country growth regressions". American Economic Review 82 (4), 942–963.
Levine, R., Zervos, S.J. (1993). "What we have learned about policy and growth from cross-country regressions?". American Economic Review 83 (2), 426–430.
Levine, R., Zervos, S. (1998). "Stock markets, banks, and economic growth". American Economic Review 88, 537–558.
Lucas, R.E. Jr. (1988). "On the mechanics of economic development". Journal of Monetary Economics 22, 3–42.
Lucas, R.E. Jr. (1990). "Why doesn't capital flow from rich to poor countries?". American Economic Review Papers and Proceedings 80 (May), 92–96.
Lucas, R.E. Jr. (1998). "The Industrial Revolution: Past and Future." Mimeo.
Maddison, A. (1995). Monitoring the World Economy, 1820–1992. Development Centre of the Organization for Economic Co-operation and Development, Paris.
Mankiw, N.G. (1995). "The growth of nations". Brookings Papers on Economic Activity 1, 275–326.
Mankiw, N.G., Romer, D., Weil, D.N. (1992). "Contribution to the empirics of economic growth". Quarterly Journal of Economics 107, 407–437.
McGrattan, E.R. (1998). "A defense of AK growth models". Quarterly Review / Federal Reserve Bank of Minneapolis 22 (Fall), 13–27.
McGrattan, E., Schmitz, J. (1998). "Explaining cross-country income differences". Mimeo, June. Federal Reserve Bank of Minneapolis, Research Department.
Murphy, K.M., Shleifer, A., Vishny, R. (1989). "Industrialization and the big push". Journal of Political Economy 97, 1003–1026.
Parente, S. (1994). "Technology adoption, learning-by-doing, and economic growth". Journal of Economic Theory 63, 346–369.
Parente, S.L., Prescott, E.C. (1996). "Barriers to technology adoption and development". Journal of Political Economy 102 (2), 298–321.
Patrinos, H.A. (1997). "Differences in education and earnings across ethnic groups in Guatemala". Quarterly Review of Economics and Finance 37 (4).
Prescott, E. (1998). "Needed: A theory of total factor productivity". International Economic Review (in press).
Pritchett, L. (1997). "Where has all the education gone?". World Bank Policy Research Working Paper No. 1581.
Pritchett, L. (1998). "Patterns of Economic Growth: Hills, Plateaus, and Mountains". World Bank Development Research Group Policy Research Working Paper No. 1947.

Pritchett, L. (1999). "The tyranny of concepts: CUDIE (cumulated, depreciated, investment effort) is not capital". Mimeo. World Bank.
Psacharopoulos, G. (1994). "Returns to investment in education: A global update". World Development 22, 1325–1343.
Psacharopoulos, G., Patrinos, H.A. (1994). "Indigenous People and Poverty in Latin America". Human Resources Development and Operations Policy Working Paper No. 22.
Quah, D. (1993). "Galton's fallacy and tests of the convergence hypothesis". Scandinavian Journal of Economics 95 (4), 427–443.
Rauch, J.E. (1993). "Productivity gains from geographic concentration of human capital: Evidence from the cities". Journal of Urban Economics 34, 380–400.
Ray, D. (1998). Development Economics. Princeton University Press, Princeton.
Ravallion, M., Jalan, J. (1996). "Growth divergence due to spatial externalities". Economics Letters 53, 227–232.
Ravallion, M., Wodon, Q. (1998). "Poor Areas or only Poor People?". Mimeo. World Bank.
Rebelo, S. (1991). "Long run policy analysis and long run growth". Journal of Political Economy 99, 500–521.
Rebelo, S. (1998). "The Role of Knowledge and Capital in Economic Development". Mimeo. Northwestern University.
Rebelo, S., Stokey, N.L. (1995). "Growth effects of flat-rate taxes". Journal of Political Economy 103, 519–550.
Rodrik, D. (1998). "Where Did All the Growth Go? External Shocks, Social Conflict and Growth Collapses". National Bureau Of Economic Research. Working Paper Series No. 6350.
Romer, P. (1986). "Increasing returns and long-run growth". Journal of Political Economy 94, 1002–1037.
Romer, P. (1990). "Endogenous technological change". Journal of Political Economy 98, S71–S102.
Romer, P. (1995). "Comment on N. Gregory Mankiw, "The growth of nations"". Brookings Papers on Economic Activity 1, 313–320.
Sachs, J., Warner, A. (1995). "Economic reform and the process of global integration". Brookings Papers on Economic Activity 1, 1–95.
Sarel, M. (1996). "Nonlinear effects of inflation on economic growth". IMF Staff Papers 43 (March), 199–215.
Shleifer, A., Vishny, R. (1993). "Corruption". Quarterly Journal of Economics 108 (3), 599–667.
Sokoloff, K.L., Engerman, S.L. (2000). "Institutions, factor endowments, and paths of development in the New World". Journal of Economic Perspectives 14 (3), 217–232.
Solow, R. (1956). "A contribution to the theory of economic growth". Quarterly Journal of Economics 70, 65–94.
Solow, R. (1957). "Technical change and the aggregate production function". Review of Economics and Statistics 39, 312–320.
Woolcock, M., Narayan, D. (2000). "Social capital: Implications for development theory, research, and policy". World Bank Research Observer 15 (2), 225–249.
World Bank (1981). Accelerated Development in Sub-Saharan Africa. Washington, DC.
Young, A. (1995). "The tyranny of numbers: Confronting the statistical realities of the East Asian growth experience". Quarterly Journal of Economics 110, 641–680.

Proof of Raw Subject Index

Proof of Raw Subject Index

Long-term Economic growth

and the History of Technology

Joel Mokyr
Departments of Economics and History
Northwestern University

Version of Oct. 2003

Prepared for the *Handbook of Economic growth,* edited by Philippe Aghion and Steven Durlauf.

Some of the material in this paper is adapted from my books *The Lever of Riches: Technological Creativity and Economic Change* New York: Oxford University Press, 1990; *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton: Princeton University Press, 2002 and *The Enlightened Economy: an Economic history of Britain, 1700-1850.* Harmondsworth: Penguin Press, 2004, as well as from a number of more detailed papers available upon request.

Introduction

As every economist knows, the modern era is the era of economic growth. In the past two centuries, measures of output per capita have increased dramatically and in a sustained manner, in a way they had never done before. It seems by now a consensus to term the start of this phenomenon "the Industrial Revolution," although it is somewhat in dispute what precisely is meant by that term (Mokyr, 1998b). In the past two decades an enormous literature has emerged to explain this phenomenon. A large number of "deep" questions have emerged which this literature has tried to answer. Below I list the most pertinent of these questions and in the subsequent pages, I shall make an attempt to answer them.

1. What explains the *location* of the Industrial Revolution (in Europe as opposed to the rest of the world, in Britain as opposed to the rest of Europe, in certain regions of Britain as opposed to others). What role did geography play in determining the main parameters of the Industrial Revolution?

2. What explains the *timing* of the Industrial Revolution in the last third of the eighteenth century (though the full swing of economic growth did not really start until after 1815)? Could it have started in the middle ages or in classical antiquity?

3. Is sustained economic growth and continuous change the "normal" state of the economy, unless it is blocked by specific "barriers to riches" or is the stationary state the normal condition, and the experience of the past 200 years is truly a revolutionary regime change?

4. What was the role of technology in the origins of the Industrial Revolution and the subsequent evolution of the more dynamic economies in which rapid growth became the norm?

5. What was the relation between demographic behavior (and specifically the fall in mortality after 1750 and the subsequent decline in fertility and shift toward fewer but higher-quality children) in bringing about and sustaining modern economic growth?

6. What was the role of institutions (in the widest sense of the word) in bringing about modern economic growth, and to what extent can we separate it from other factors such as technology and factor accumulation?

7. To what extent is modern growth due to "culture," that is, intellectual factors regarding beliefs, attitudes, and preferences? Does culture normally adapt to the economic environment, or can one discern autonomous cultural changes that shaped the economy?

8. Did the "Great Divergence" really start only in the eighteenth century, and until then the economic performance and potential of occident and the orient were comparable, or can signs of the divergence be dated to the renaissance or even the middle ages?

9. Was the Industrial Revolution "inevitable" in the sense that the economies a thousand years earlier already contained the seeds of modern economic growth that inexorably had to sprout and bring it about?

10. What was the exact role of human capital, through formal education or other forms, in bringing about modern economic growth?

Technology and Economic growth

Economists have become accustomed to associate long-term economic growth with technological progress; it is deeply embedded in the

main message of the Solow-inspired growth models, which treated technological change as exogenous, and even more so in the endogenous growth models.[1] Whether technology is a deus ex machina that somehow makes productivity grow a little each year, or produced within the system by the rational and purposeful application of research and development, the growth of human and physical capital that are strongly complementary with productivity growth, or even in the simple TFP computations that often equate the residual with technological progress — technology is central to the dynamic of the economy in the past two centuries. Many scholars believe that people are inherently innovative and that if only the circumstances are right (the exact nature of these conditions differs from scholar to scholar), technological progress is almost guaranteed. This somewhat heroic assumption is shared by scholars as diverse as Robert Lucas and Eric L. Jones, yet it seems at variance with the historical record. The record is that despite many significant, even path-breaking innovations in many societies since the start of written history, it has not really been a major factor in economic growth, such as it was, before the Industrial Revolution.

Instead, economic historians studying earlier periods have come to realize that technology was less important than institutional change in explaining pre-modern episodes of economic growth. It is an easy exercise to point to the many virtues of "Smithian Growth," the increase in economic output due to commercial progress (as opposed to technological progress). Better markets, in which agents could specialize according to their comparative advantage and take full advantage of economies of scale, and in which enhanced competition would stimulate efficiency and the adoption of best-practice technology could generate growth sustainable for decades and even centuries. Even with no changes whatsoever in technology, economies can grow in the presence of peace, law and order, improved communications and trust, the introduction of money and credit, enforceable and secure property rights, and similar institutional improvements (Greif, 2003). Similarly, better institutions can lead to improved

[1] The opening line of the standard textbook in the area states that the "most basic proposition of growth theory is that in order to sustain a positive growth rate of output per capita in the long run, there must be continual advances in technological knowledge" (Aghion and Howitt, 1998, p. 11).

allocation of resources: law and order and improved security can and will encourage productive investment, reduce the waste of talent on rent-seeking and the manipulation of power for the purposes of redistribution (North, 1990; Shleifer and Vishny, 1998; Baumol, 2002). Tolerance for productive "service minorities" who lubricated the wheels of commerce (Syrians, Jews and many others) played important roles in the emergence of commerce and credit. Economic history before 1750 is primarily about this kind of growth. The wealth of Imperial Rome and the flourishing of the medieval Italian and Flemish cities, to pick just a few examples, were based above all on commercial progress, sometimes referred to as "Smithian Growth."

It is usually assumed by economists that sustained economic growth is a recent phenomenon simply because if modern rates of growth had been sustained, a simple backward projection suggests that income in 1500 or in 1000 would have been absurdly low.[2] Clearly, growth at the rates we have gotten used to in the twentieth century are unthinkable in the long run. Yet it is equally implausible to think that just because growth was slower, there was *none* of it – after all, there is a lot of time in the long run. One does not have to fully subscribe to Graeme Snooks's use of Domesday book and Gregory King's numbers 600 years later to accept his view that by 1688 the British economy was very different indeed from what it had been at the time of William the Conqueror. Adam Smith had no doubt that "the annual produce of the land and labour of England... is certainly much greater than it was a little more than century ago at the restoration of Charles II (1660)... and [it] was certainly much greater at the restoration than we can suppose it to have been a hundred years before" (Smith, 1776-1976, pp. 365-66).[3] On the eve of the Industrial Revolution, large parts of

---

[2] For instance, income per capita in the UK in 1890 was about $4100 in 1990 international dollars. It grew in the subsequent years by an average of 1.4% per year. Had it been growing at that same rate in the previous 300 years, income percapita in 1590 would have been $ 61, which clearly seems absurdly low.

[3] Snooks's (1994) belief in pre-modern growth is based essentially on his comparison between the income per capita he has calculated from the Domesday book (1086) and the numbers provided by Gregory King for 1688. While such computations are of course always somewhat worrisome (what, exactly, does it mean to estimate the

Europe and some parts of Asia were enjoying a standard of living that had not been experienced ever before, in terms of the quantity, quality, and variety of consumption.[4] Pre-1750 growth was primarily based on Smithian and Northian effects: gains from trade and more efficient allocations due to institutional changes. The Industrial Revolution, then, can be regarded not as the beginnings of growth altogether but as the time at which technology assumed an ever-increasing weight in the generation of growth.[5] It should not be confused with the demographic transition, which came later and whose relationship with technological progress is complex and poorly understood.

This is not to say that before the Industrial Revolution technology was altogether unimportant in its impact on growth. Medieval Europe was an innovative society which invented many important things (including movable type, gunpowder, spectacles, the mechanical clock) and adopted many more inventions from other societies (paper, navigational instruments, Arabic numerals, the lateen sail, wind power). Yet, when all is said and done, it is hard to argue that the impact of these inventions on the growth of GDP or some other measure of aggregate output were all that large. The majority of the labor force was still employed in agriculture where progress was exceedingly slow (even if over the long centuries be-

---

nominal income of 1086 in the prices of 1688 given the many changes in consumption items?), the order of magnitude provided by Snooks (an increase of real income by 580 percent) may survive such concerns. Medievalists tend to agree with the occurrence of economic growth in Britain, though their figures indicate a much slower rate of growth, about a 111 percent growth rate between 1086 and 1470 (Britnell, 1996, p. 229), which would require more economic growth in the sixteenth and seventeenth centuries than can be justified to square with Snooks's numbers. Engerman (1994, p. 116) assesses that most observers will agree with Snooks's view that by 1700 England had a high level of per capita income and was in a good position to "seek the next stage of economic growth." Yet clearly he is correct in judging that "modern" economic growth (prolonged, continuous, rapid) did not begin until the early nineteenth century.

[4] Indeed, many historians speak of a "consumer revolution" *prior to* the Industrial Revolution, which would be inexplicable without rising income before 1750. Lorna Weatherill (1988) suggests that if there was a Consumer Revolution at all, it peaked in the period 1680-1720. Moreover, consumer revolutions were taking place elsewhere in Europe. Seventeenth century Holland was, of course, the most obvious example thereof, but Cissie Fairchilds (1992) has employed probate records to show that France, like England, experienced a consumer revolution, albeit fifty years later.

[5] It is in that sense that the view of modern economists (e.g. Galor and Weil, 2000, p. 809) that "the key event that separates Malthusian and Post-Malthusian regimes is the acceleration of the pace of technological progress" is a bit misleading, since it draws a link between technological progress and demographic change that thus far has not been closely examined.

tween 800 and 1300 the three-field system and the growing efficiency at which livestock was employed did produce considerable productivity gains).

Moreover, it is true for the pre-1750 era – as it was a fortiori after 1750 – that technology itself interacted with Smithian growth because on balance improved technology made the expansion of trade possible – above all maritime technology in all its many facets, but also better transport over land and rivers, better military technology to defeat pirates, better knowledge of remote lands, and the growing ability to communicate with strangers. A decomposition of growth into a technology component and a trade-and-institutions component must take into account such interactions.

All the same, the main reason why technological progress was at best an also-ran in the explanation of economic growth before 1750 is that even the best and brightest mechanics, farmers, and chemists — to pick three examples — knew relatively little of what could be known about the fields of knowledge they sought to apply. The pre-1750 world produced, and produced well. It made many pathbreaking inventions. But it was a world of engineering without mechanics, iron-making without metallurgy, farming without soil science, mining without geology, water-power without hydraulics, dye-making without organic chemistry, and medical practice without microbiology and immunology. Not enough was known to generate sustained economic growth based on technological change. Such statements are of course to some extent provocative and perhaps even irresponsible: how can we define "what could be known" in any meaningful sense? Who knew "that which was known" and how did they use it? In what follows I shall propose a simple framework to understand how and why new technology emerges and how it was limited before the eighteenth century and then liberated from its constraints. I will then argue that "technological modernity" means an economy in which *sustained* technological progress is the primary engine of growth and that it depended on the persistence of technological progress.

A Historical Theory of Technology

Technology is knowledge. Knowledge, as is well known, has always been a difficult concept for standard economics to handle. It is at the core of modern economic growth, but many characteristics make it slippery to handle. Knowledge is above all a non-rivalrous good, that is, sharing it with another person does not diminish the knowledge of the original owner. It is not quite non-excludable, but clearly excludability is costly and for many types of knowledge exclusion costs are infinite. It is produced in the system, but the motivation of its producers are rarely purely economic. Indeed, the producers of scientific knowledge almost never collect but a

tiny fraction of the surplus they produce for society. It is the mother of all spillover effects. A more fruitful approach than to view knowledge as an odd sort of good, pioneered by Olsson (2000, 2003), is to model knowledge as a set, and to analyze its growth in terms of the properties of existing knowledge rather than looking at the motivations of individual agents.

The basic unit of analysis of technology is the "technique." A technique is a set of instructions, much like a cookbook recipe, on how to produce goods and services. As such, it is better defined than the concept of a stock of "ideas" that some scholars prefer (e.g. Charles Jones, 2001). The entire set of feasible techniques that each society has at its disposal is bound by the isoquant. Each point on or above the isoquant in principle represents a set of instructions on how to combine various ingredients in some way to produce a good or service that society wants. While technology often depends on artifacts, the artifacts are not the same as the technique and what defines the technique is the content of the instructions. Thus, a piano is an artifact, but what is done with it depends on the technique used by the pianist, the tuner, or the movers. Society's production possibilities are bound by what society knows. This knowledge includes the knowledge of making artefacts and using them.

But who is "society"? The only sensible way of defining knowledge at a social level is as the *union* of all the sets of individual knowledge. This definition is consistent with our intuitive notion of the concept of an invention or a discovery – at first only *one* person has it, but once that happens society as a whole feels it has acquired it. Knowledge can be stored in external storage devices such as books, drawings, and artifacts but such knowledge is meaningless unless it can be transferred to an actual person. Such a definition immediately requires a further elaboration: if one person possesses a certain knowledge, how costly is it for others to acquire it? This question, indeed is at the heart of the idea of a "technological society." Knowledge is shared and distributed, and its transmission through learning is essential for such a society to make effective use of it. Between the two extremes of a society in which all knowledge acquired by one member is "episodic" and not communicated to any other member, and the other extreme in which all knowledge is shared through some monstrous network, there was a reality of partial and costly sharing and access. But these costs were not historically invariant, and their changes are one of the keys to technological change.

Progress in exploiting the existing stock of knowledge will depend first and foremost on the efficiency and cost of access to knowledge. Although knowledge is a public good in the sense that the consumption of one does not reduce that of others, the private costs of acquiring it are not negligible, in terms of time, effort, and often other real resources as well (Reiter, 1992, p. 3). Access costs include the costs of finding out whether an

answer to a question actually exists, if so, where it can be found, then paying the cost of acquiring it, and finally verifying the correctness of the knowledge. When the access costs become very high, it could be said in the limit that social knowledge has disappeared.[6] Language, mathematical symbols, diagrams, and physical models are all means of reducing access costs. Shared symbols may not always correspond with the things they signify, as postmodern critics believe, but as long as they are shared they reduce the costs of accessing knowledge held by another person or storage device.

The determinants of these access costs are both institutional and technological: "open knowledge" societies, in which new discoveries are published as soon as they are made and in which new inventions are placed in the public domain through the patenting system (even if their application may be legally restricted), are societies in which access costs will be lower than in societies in which the knowledge is kept secret or confined to a small and closed group of insiders whether they are priests, philosophers, or mandarins. Economies that enjoyed a high level of commerce and mobility were subject to knowledge through the migration of skilled workmen and the opportunities to imitate and reverse-engineer new techniques. As access costs fell in the early modern period, it became more difficult to maintain intellectual property rights through high access costs, and new institutions that provided incentives for innovators became necessary, above all the patent system emerging in the late fifteenth and sixteenth centuries. The printing press clearly was one of the most

---

[6] This cost function determines how costly it is for an individual to access information from a storage device or from another individual. The *average* access cost would be the average cost paid by all individuals who wish to acquire the knowledge. More relevant for most useful questions is the *marginal* access cost, that is, the *minimum* cost for an individual who does not yet have this information. A moment reflection will make clear why this is so: it is very expensive for the average member of a society to have access to the Schrödinger wave equations, yet it is "accessible" at low cost for advanced students of quantum mechanics. If someone "needs" to know something, he or she will go to an expert for whom this cost is as low as possible to find out. Much of the way knowledge has been used in recent times has relied on such experts. The cost of finding them experts and retrieving knowledge thus determines marginal access costs. Equally important, as we shall see, is the technology that provides access to storage devices.

significant access-cost-reducing inventions of the historical past.[7] The nature of the books printed, such as topic, language, and accessibility, played an equally central role in their reduction. People normally acquired knowledge and skills vertically, but also from one another through imitation. Postdoctoral students in laboratory settings full-well realize the differences between the acquisition of codifiable knowledge and the acquisition of tacit knowledge through imitation and a certain *je ne sais quoi* we call experience.[8]

Techniques constitute what I have called *prescriptive* knowledge – like any recipe they essentially comprise instructions that allow people to "produce," that is, to exploit natural phenomena and regularities in order to improve human material welfare.[9] The fundamental unit of set of prescriptive knowledge has the form of a set of do-loops (often of great complexity, with many if-then statements), describing the "hows" of what we call production.

There are two preliminary observations we need to point out in this context. One is that it is impossible to specify explicitly the entire

---

[7] Elizabeth Eisenstein (1979) has argued that the advent of printing created the background on which the progress of science and technology rests. In her view, printing created a "bridge over the gap between town and gown" as early as the sixteenth century, and while she concedes that "the effect of early printed technical literature on science and technology is open to question" she still contends that print made it possible to publicize "socially useful techniques" (pp. 558, 559).

[8] It should be obvious that in order to read such a set of instructions, readers need a "codebook" that explains the terms used in the technique (Cowan and Foray, 1997). Even when the techniques are explicit, the codebook may not be, and the codebook needed to decipher the first codebook and the next, and so on, eventually must be tacit. Sometimes instructions are "tacit" even when they could be made explicit but it is not cost-effective to do so.

[9] These instructions are essentially identical to the concept of "routines|" proposed by Nelson and Winter (1982). When these instructions are carried out in practice, we call it production, and then they are no longer knowledge but action. "Production" here should be taken to include household activities such as cooking, cleaning, childcare, and so forth, which equally require the manipulation of natural phenomena and regularities. It is comparable to DNA instructions being "expressed." Much like instructions in DNA, the lines in the technique can be either "obligate" (do X) or "facultative" (if Y, do X). For more complex techniques, nested instructions are the rule.

content of a set of instructions. Even a simple cooking recipe contains a great deal of assumptions that the person executing the technique is supposed to know: how much a cup is, when water is boiling, and so on. For that reason, the person executing a technique is supposed to have certain knowledge that I shall call *competence* to distinguish it from the knowledge involved in writing the instructions for the first time (that is, actually making the invention). Competence consists of the knowledge of how to read, interpret, and execute the instructions in the technique and the supplemental tacit knowledge that cannot be fully written down in the technique's codified instructions. There is a continuum between the implicit understandings and clever tricks that make a technique work we call tacit knowledge, and the minor improvements and refinements introduced subsequent to invention that involve actual adjustments in the explicit instructions. The latter would be more properly thought off as microinventions, but clearly any sharp distinction between them would be arbitrary. All the same, "competence" and "knowledge" are no less different than the differences in skills needed to play the Hammerklavier sonata and those needed to compose it. One of the most interesting variables to observe is the ratio between the knowledge that goes into the first formulation of the technique in question (invention) and the competence needed to actually carry out the technique. As we shall see, it is this ratio around which the importance of human capital in economic growth will pivot.

The second observation is the notion that every technique, because it involves the manipulation and harnessing of natural regularities, requires an *epistemic base,* that is, a knowledge of nature on which it is based. I will call this type of knowledge *propositional* knowledge, since it contains a set of propositions about the physical world. The distinction between propositional and prescriptive knowledge seems obvious: the planet Neptune and the structure of DNA were not "invented"; they were already there prior to discovery, whether we knew it or not. The same cannot be said about diesel engines or aspartame. Polanyi notes that the distinction is recognized by patent law, which permits the patenting of inventions (additions to prescriptive knowledge) but not of discoveries (additions to propositional knowledge). He points out that the difference boils down to observing that prescriptive knowledge can be "right or wrong" whereas "action can only be successful or unsuccessful."(1962, p. 175). Purists will object that "right" and "wrong" are judgments based on socially constructed criteria, and that

"successful" needs to be defined in a context, depending on the objective function that is being maximized.

The two sets of propositional and prescriptive knowledge together form the set of useful knowledge in society. These sets satisfy the conditions set out by Olsson (2000) for his "idea space." Specifically, the sets are infinite, closed, and bounded. They also are subsets of much larger sets, the sets of knoweable knowledge. At each point of time, the actual sets describe what a society knows and consequently what it can do. There also is a more complex set of characteristics that connect the knowledge at time t with that in the next period. Knowledge is mostly cumulative and evolutionary. The "mostly" is added because it is not wholly cumulative (knowledge *can* be lost, though this has become increasingly rare) and its evolutionary features are more complex than can be dealt with here (Mokyr, 2003a).

The actual relation between propositional and prescriptive knowledge can be summarized in the following 10 generalizations:

1. Every technique has a minimum epistemic base, which contains the least knowledge that society needs to possess for this technique to be invented. The epistemic base contains at the very least the trivial statement that technique i works.[10] There are and have been some techniques, invented accidentally or through trial and error, about whose modus operandi next to nothing was known except that they worked. We can call these techniques *singleton* techniques (since their domain is a singleton).

2. Some techniques require a minimum epistemic base larger than a singleton for a working technique to emerge. It is hard to imagine such techniques as nuclear resonance imaging or computer assisted design software as emerging in any society as the result of serendipitous finds or trial-and-error methods, without the designers having a clue of why and how they worked.

[10] This statement is true because the set of propositional knowledge contains as a subset the list (or catalog) of the techniques that work – since that statement can be defined as a natural regularity.

3. The actual epistemic base is equal to or larger than the minimum epistemic base. It is never bound from above in the sense that the amount that can be known about the natural phenomena that govern a technique is infinite. In a certain sense, we can view the epistemic base at any given time much like a fixed factor in a production function. As long as it does not change, it imposes concavity and possibly even an upper bound on innovation and improvement. On the other hand, beyond a certain point, the incremental effect of widening the actual epistemic base on the productivity growth of a given technique will run into diminishing returns and eventually be limited.

4. There is no requirement that the epistemic base be "true" or "correct" in any sense. In any event, the only significance of such a statement would be that it conforms to contemporary beliefs about nature (which may well be refuted by future generations). Thus the humoral theory of disease, now generally rejected, formed the epistemic base of medical techniques for many centuries. At the same time, some epistemic bases can be more effective than others in the sense that techniques based on them perform "better" by some agree-upon criterion. "Effective knowledge" does not mean "true" – many techniques were based on knowledge we no longer accept yet were deployed for long periods with considerable success.[11]

5. The wider the actual epistemic base supporting a technique relative to the minimum one, the more likely an invention is to occur, ceteris paribus. A wider epistemic base means that it is less likely for a researcher to enter a blind alley and to spend resources

[11] Here one can cite many examples. Two of them are the metallurgical writings and inventions of René Réaumur and Tobern Bergman, firmly based on phlogiston physics, and the draining of swamps based on the belief that the "bad air" they produced caused malaria.

in trying to create something that cannot work.[12] Thus, a wider epistemic base reduces the costs of research and development and increases the likelihood of success.

6. The wider the epistemic base, the more likely an existing technique is to be improved, adapted, and refined through subsequent microinventions. The more is known about the principles of a technique, the lower will be the costs of development and improvement. This is above all because the more is known *why* something works, the better the inventor can tweak its parameters to optimize and debug the technique. Furthermore, because invention so often consists of analogy with or the recombination of existing techniques, lower access cost to the catalog of existing techniques (which is part of propositional knowledge) stimulates and streamlines successful invention.

7. The epistemic bases in existence during the early stages of an invention are historically usually quite narrow at first, but are often enlarged following the appearance of the invention, and sometimes directly on account of the invention.

8. Both propositional and prescriptive knowledge can be "tight" or "untight." Tightness measures the degree of confidence and consensualness of a piece of knowledge: how sure are people that the knowledge is "true" or that the technique "works?" The tighter a piece of propositional knowledge, the lower are the costs of verification and the more likely the technique is to be adopted, and vice versa. Of course, tightness should be closely correlated with effectiveness: a laser printer works better than a dot matrix, and

---

[12] Alchemy – the attempt to turn base metals into gold by chemical means – was still a major occupation of the best minds of the scientific revolution above all Isaac Newton. By 1780 Alchemy was in sharp decline and in the nineteenth century chemists knew enough to realize that it was a misallocation of human capital to search for the stone of the wise as it was for the fountain of youth. The survival of astrology in our time demonstrates that the prediction of the future – always a technique based on a very narrow epistemic base – has not benefitted in a similar way from a widening of the prescriptive knowledge on which it was based.

there can be little dispute about the characteristics here. If two techniques are based on incompatible epistemic bases, the one that works better will be chosen and the knowledge on which it is based will be judged to be more effective. But for much of history, such testing turned out to be difficult to do and propositional knowledge was more often selected on the basis of authority and tradition that effectiveness. Even today, for many medical and farming techniques it is often difficult to observe what works and what does not work as well without careful statistical analysis or experimentation.

9. It is not essential that the person writing the instructions actually knows himself everything that is in the epistemic base. Even if very few individuals in a society know quantum mechanics, the practical fruits of the insights of this knowledge to technology may still be available just as if everyone had been taught advanced physics. It is a fortiori true that the people carrying out a set of instructions do not know how and why these instructions work, and what the support for them is in propositional knowledge. No doctor prescribing nor any patient taking an aspirin will need to study the biochemical properties of prostaglandins, though such knowledge may be essential for those people trying to design an analgesic with, say, fewer side effects. What counts is collective knowledge and the cost of access as discussed above. It is even less necessary for the people actually carrying out the technique to possess the knowledge on which it is based, and normally this is not the case.

10. The existence of a minimum epistemic base is a necessary but insufficient condition for a technique to emerge. A society may well accumulate a great deal of propositional knowledge that is never translated into new and improved techniques. Knowledge opens doors, but it does not force society to walk through them.

The significance of the Industrial Revolution.

Historians in the 1990s have tended to belittle the significance of the Industrial Revolution as a historical phenomenon, referring to it as the so-called Industrial Revolution, and pointing to the slowness and gradualness of economic change, as well as the many continuities that post 1760 Britain had with earlier times (for a critical survey, see Mokyr 1998b).

Before I get to the heart of the argument, two points need to be cleared away. The first is the myth that the Industrial Revolution was a purely British affair, and that without Britain's leadership Europe today would still be largely a subsistence economy. The historical reality was that many if not most of the technological elements of the Industrial Revolution were the result of a joint international effort in which French, German, Scandinavian, Italian, American and other "western" innovators collaborated, swapped knowledge, corresponded, met one another, and read each others' work.

It is of course true that in most cases the first successful economic *applications* of the new technology appeared in Britain. Clearly in 1789 Britain had an advantage in the execution of new techniques. Yet an overwhelming British advantage in *inventing* — especially in generating the crucial macroinventions that opened the doors to a sustained trajectory of continuing technological change — is much more doubtful, and their advantage in expanding the propositional knowledge that was eventually to widen the epistemic bases of the new techniques is even more questionable. Britain's technological precociousness in the era of the Industrial Revolution was a function of three factors.

First, it was at peace in a period when the Continent was engulfed in political and military upheaval. Not only that there was no fighting on British soil; the French revolution and the Napoleonic era was a massive distraction of talent and initiative that would otherwise have been available to technology and industry.[13] The attention of both decision makers and

---

[13] The great chemists Claude Berthollet and Jean-Antoine Chaptal, for instance, both directed their abilities to administration during the Empire. Their illustrious teacher, the great Lavoisier himself, was executed as a tax farmer. Another example is Nicolas de Barneville, who was active in introducing British spinning equipment into France. De Barneville repeatedly was called upon to serve in military positions and was "one of those unfortunate individuals whose lives have been marred by war and revolution ... clearly a victim of the troubled times" (McCloy, 1952, pp. 92-94).

inventors was directed elsewhere.[14] During the stormy years of the Revolution, French machine breakers found an opportunity to mount an effective campaign against British machine, thus delaying their adoption (Horn, 2003). Second, Britain's entrepreneurs proved uncannily willing and able to adopt new inventions regardless of where they were made, free from the "not made here" mentality of other societies. Some of the most remarkable inventions made on the Continent were first applied on a wide scale in Britain. Among those, the most remarkable were gas-lighting, chlorine bleaching, the Jacquard loom, the Robert continuous paper-making machine, and the Leblanc soda making process.[15] In smaller industries, too, the debt of the British Industrial Revolution to Continental technology demonstrates that in no sense did Britain monopolize the inventive process.[16] The British advantage in application must be chalked up largely

---

[14] The Frenchman Philippe LeBon, co-inventor of gaslighting in the 1790s, lost out in his race for priority with William Murdoch, the ingenious Boulton and Watt engineer whose work in the end led the introduction of this revolutionary technique in the illumination of the Soho works in 1802. As one French historian sighs, "during the terrors of the Revolution... no one thought of street lights. When the mob dreamed of lanterns, it was with a rather different object in view" (Cited by Griffiths, 1992, p. 242).

[15] Nicholas Leblanc, who developed the soda making process named after him. Leblanc reacted salt and sulphuric acid to produce sodium sulphate, which after heating with lime or charcoal yielded raw soda together with hydrochloric acid, a noxious by-product. The Leblanc process became the basis of the modern chemical industry and is regarded as one of the most important inventions of the time. In the adoption of soda, Britain was relatively slow, and only in the 1820s did it start to adopt Leblanc's process on a large scale. The explanation usually given for this delay is the high tax on salt, which made artificial soda more expensive than vegetable alkali. Once the salt tax was repealed, British soda production grew rapidly and by the 1850s exceeded French output by a factor of three (Haber, 1958, pp. 10-14).

[16] The most important breakthrough in the glass industry was made in 1798 by Pierre Louis Guinand, a Swiss, who invented the stirring process in which he stirred the molten glass in the crucible using a hollow cylinder of burnt fireclay, dispersing the air bubbles in the glass more evenly. The technique produced optical glass of unprecedented quality. Guinand kept his process secret, but his son sold the technique to a French manufacturer in 1827, who in turn sold it to the Chance Brothers Glass Company in Birmingham, which soon became one of the premier glassmakers in Europe. The idea of preserving food by cooking followed by vacuum sealing was hit upon by the Frenchman Nicolas Appert in 1795. Appert originally used glassware to store preserved foods, but in 1812 an Englishman named Peter Durand suggested using

to its comparative advantage in microinventions and in the supply of the human capital that could carry out the new techniques.[17] To employ the terminology proposed earlier: Britain may not have more propositional knowledge available for its invention and innovation process, but if it workers possessed higher levels of competence, then the new techniques that had emerged would find their first applications there. Its system of informal technical training, through master-apprentice relationships, created workers of uncommon skill and mechanical ability. This system produced, of course, inventors: the most famous of these such as the clockmakers John Harrison and Benjamin Huntsman, the engineer John Smeaton, the instrument maker Jesse Ramsden, the wondrously versatile inventor Richard Roberts, the chemists James Keir and Joseph Black, and of course the great Watt himself were only the first row of a veritable army of people, who in addition to possessing formal knowledge, were blessed by a technical intuition and dexterity reaching into the deeps of tacit knowledge.

Third, by the middle of the eighteenth century Britain had developed an institutional strength and agility that provided it with a considerable if temporary advantage over its Continental competitors: it had a healthier public finance system, weaker guilds, no internal tariff barriers, a superior internal transportation system, fairly well-defined and enforced property rights on land (enhanced by Parliamentary acts when necessary), and a power structure that favored the rich and the propertied classes. Moreover, it had that most elusive yet decisive institutional feature that makes for economic success: the flexibility to adapt its economic and legal institutions without political violence and disruptions. Britain's great asset was not so much that she had "better" government but rather that its

---

tin-plated cans, which were soon found to be superior. By 1814, Bryan Donkin was supplying canned soups and meats to the Royal Navy.

[17] This was already pointed by Daniel Defoe, who pointed out in 1726 that "the English ... are justly fam'd for improving Arts rather than inventing" and elsewhere in his Plan of English Commerce that "our great Advances in Arts, in Trade, in Government and in almost all the great Things we are now Masters of and in which we so much exceed all our Neighbouring Nations, are really founded upon the inventions of others." The great engineer John Farey, who wrote an important treatise on steam power, testified a century later that "the prevailing talent of English and Scotch people is to apply new ideas to use, and to bring such applications to perfection, but they do not imagine as much as foreigners."

political institutions were nimbler, and that they could be changed at low social cost by a body assigned to changing the rules and laws by which the economic game was played. Many of the rules still on the books in the eighteenth century were not enforced, and rent seeking arrangements, by comparison, were costly to attain and uncertain in their yield. British mercantilist policy was already in decline on the eve of the Industrial Revolution. Yet as the Industrial Revolution unfolded, it required further change in the institutional basis of business. The Hanoverian governments in Britain were venal and nepotist, and much of the business of government was intended to enrich politicians. But with the growing notion that rent seeking was harmful, this kind of corruption weakened. As Porter (1990, p. 119) put it, with the rise of the laissez faire lobby, Westminster abandoned its long-standing mercantilist paternalism, repealing one regulation after another. Abuses may have been deep rooted, and rent-seekers resisted all they could, but form the last third of the eighteenth century on rent-seeking was on the defensive, and by 1835 many of the old institutions had vanished, and the British state, for a few decades, gave up on redistributing income as a main policy objective. Following North (1990, p. 80) we might call this adaptive efficiency, meaning not only the adaptation of the allocation of resources but of the institutions themselves. To bring this about, what was needed was a meta-institution such as parliament that was authorized to change the rules in a consensual manner.

Compared to Britain, the Continental countries had to make a greater effort to cleanse their economic institutions from medieval debris and the fiscal ravages of absolutism, undo a more complex and pervasive system of rent seeking and regulation, and while extensive reforms were carried out in France, Germany, and the Low Countries after the French Revolution, by 1815 the work was still far from complete and had already incurred enormous social costs. It took another full generation for the Continent to pull even. None of the British advantages was especially deep or permanent. They explain Britain's position as the lead car in the Occident Express that gathered steam in the nineteenth century and drove away from the rest of the world, but it does not tell us much about the source of power. Was Britain the engine that pulled the other European cars behind it, or was Western Europe on an electric train deriving its motive power from a shared source of motive energy?

One useful mental experiment is to ask whether there would have been an Industrial Revolution in the absence of Britain. A counterfactual industrial revolution led by Continental economies would have been delayed by a few decades and differed in some important details. It might have relied less on "British" steam and more on "French" water power and "Dutch" wind power technology, less on cotton and possibly more on wool and linen. It would have more of an *étatist* flavor, with a bigger emphasis on military engineering and public projects. But in view of the capabilities of French engineers and German chemists, the entrepreneurial instincts of Swiss and Belgian industrialists, and the removal of many institutions that had hampered their effective deployment before 1789, a technological revolution not all that different from what actually transpired would have happened. Even without Britain, by the twentieth century the gap between Europe and the rest of the world would have been there (Mokyr, 2000).

The second point to note is that the pivotal element of the Industrial Revolution took place later than is usually thought. The difference between the Industrial Revolution of the eighteenth century and other episodes of a clustering of macroinventions was not just in the celebrated inventions in the period 1765-1790. While the impact of the technological breakthroughs of these years of *sturm und drang* on a number of critical industries stands undiminished, the critical difference between this Industrial Revolution and previous clusters of macroinventions is not that these breakthroughs occurred at all, but that their momentum did not level off and peter out after 1800 or so. In other words, what made the Industrial Revolution into the "great divergence" was the *persistence* of technological change after the first wave. We might well imagine a counterfactual technological steady state of throstles, wrought iron, and stationary steam engines, in which there was a one-off shift from wool to cotton, from animate power to stationary engines, and of cheap wrought iron. It is easy to envisage the economies of the West settling into these techniques without taking them much further, as had happened in the wave of inventions of the fifteenth century.

But this is not what happened. The "first wave" of innovations was followed after 1820 by a secondary ripple of inventions that may have been less spectacular, but these were the microinventions that provided the muscle to the downward trend in production costs. The second stage of the Industrial Revolution adapted novel ideas and tricks to be applied in new

and more industries and sectors, improved and refined the earlier and eventually showed up in the productivity statistics. Among those we may list the perfection of mechanical weaving after 1820; the invention of Roberts's self-acting mule in spinning (1825); the extension and adaptation of the techniques first used in cotton to carded wool and linen; the continuing improvement in the iron industry through Neilson's hot blast (1829) and other inventions; the continuing improvement in steampower, raising the efficiency and capabilities of the low pressure stationary engines, while perfecting the high pressure engines of Trevithick, Woolf, and Stephenson and adapting them to transportation; the advances in chemicals before the advent of organic chemistry (such as the breakthroughs in candle-making and soap manufacturing thanks to the work of Eugène-Michel Chevreul on fatty acids); the introduction and perfection of gas-lighting; the breakthroughs in engineering and high-precision tools by Maudslay, Whitworth, Nasmyth, Rennie, the Brunels, the Stephensons, and the other great engineers of the "second generation"; the growing interest in electrical phenomena leading to electroplating and the work by Hans Oersted and Joseph Henry establishing the connection between electricity and magnetism, leading to the telegraph in the late 1830s; the continuous improvement in crucible steelmaking through coordinated crucibles (as practiced for example by Krupp in Essen); the pre-Bessemer improvements in steel thanks to the work of Scottish steelmakers such as David Mushet (father of Robert Mushet, celebrated in one of Samuel Smiles's *Industrial Biographies*), and the addition of manganese to crucible steel known as Heath's process (1839).

The second wave of inventions was the critical period in the sense that it shows up clearly in the total income statistics. Income per capita growth after 1830 accelerates to around 1.1 percent, even though recent calculations confirm that only about a third of that growth to total factor productivity growth (Antras and Voth, 2003, p. 63; Mokyr, 2003c). Income growth in Britain during the "classical" Industrial Revolution was modest. This fact is less difficult to explain than some scholars make it out to be, and any dismissal of the Industrial Revolution as a historical watershed for that reason seems unwarranted. After all, the disruptions of international commerce during the quarter century of the French Wars coincided with bad harvests and unprecedented population growth. Yet the main reason is simply that in the early decades the proportion of the British economy

affected by technological progress and that can be viewed as a "modern sector" was simply small, even if its exact dimensions remain in dispute. From 1830 this sector expands rapidly as the new technology is applied more broadly (especially to transportation), growth accelerates, and by the mid 1840's there is clear-cut evidence that the standard of living in Britain is rising even for the working class. It also serves as a bridge between the first Industrial Revolution and the more intense and equally dramatic changes of the second Industrial Revolution I will discuss below.

To sum up, then, the period 1760-1830 Western Europe witnessed a growing importance of invention, the emergence of new techniques that in the longer run were to have an enormous impact on productivity and growth. Without belittling the other elements that made the Industrial Revolution possible, the technological breakthroughs of the period prepared the ground for the economic transformation that made the difference between the West and the Rest, between technological modernity and the much slower and often-reversed economic growth episodes of the previous millennia. In order to come up with a reasonable explanation of the technological roots of economic growth in this period, we must turn to the intellectual foundations of the explosion of technical knowledge.

The Intellectual Roots of the Industrial Revolution

Economic historians like to explain economic phenomena with other economic phenomena. The Industrial Revolution, it was felt for many decades should be explained by economic factors. Relative prices, better property rights, endowments, changes in fiscal and monetary institutions, investment, savings, exports, and changes in labor supply have all been put forward as possible explanations (for a full survey, see Mokyr, 1998). Yet the essence of the Industrial Revolution was technological, and technology is knowledge. How, then, should we explain not just the famous inventions of the Industrial Revolution but also the equally portentous fact that these inventions did not peter out fairly quickly after they emerged, as had happened so often in the past?

The answer has to be sought in the intellectual changes that occurred in Europe *before* the Industrial Revolution. These changes affected the sphere of propositional knowledge. The problem, as economic historians have known for many years, is that it is very difficult to argue that the

scientific revolution of the seventeenth century we associate with Galileo, Descartes, Newton, and many other giants, had a direct impact on the Industrial Revolution (McKendrick, 1973). Few important inventions, both before and after 1800, can be directly attributed to great scientific discoveries or were dependent in any direct way on scientific expertise. The advances in physics, chemistry, biology, medicine, and other areas occurred too late to have the desired effect. The scientific advances of the seventeenth century, crucial as they were to the understanding of nature, had more to do with the movement of heavenly bodies, optics, magnetism, and the classification of plants than with the motions of machines. To say that therefore it had no economic significance is an exaggeration: many of the great scientists wrote about mechanics and the properties of materials.[18] Yet it is hard to see many examples of eighteenth-century inventions that owed their existence to a prior scientific discovery.[19] After 1800 the

---

[18] From the viewpoint of the history of technology, Galileo is particularly important because his theory of mechanics and concept of force lies at the basis of all machines. Until Galileo, the idea that general laws governed all machines was not recognized; each machine was described as if it were unique. Galileo realized that all machines transmitted and applied force as special cases of the lever and fulcrum principle. As Cardwell points out, Galileo's theory of mechanics is interesting to the economist because the concept governing it is one of efficiency: "The function of a machine is to deploy and use the powers that nature makes available in the best possible way for man's purposes... the criterion is the amount of work done --- however that is evaluated --- and not a subjective assessment of the effort put into accomplishing it" (Cardwell, 1972, pp. 38-39). In the writings of Galileo, the leading scientist of his time, economic efficiency is linked with science. In his *Motion and Mechanics* he wrote that the advantage of machines was to harness cheap sources of energy because "the fall of a river costs little or nothing." In this he differed radically from his inspiration, Archimedes, and this difference between the two scientific giants who established the science of mechanics epitomizes the difference between classical and early modern society. The great French physicist René Réaumur (1683-1757) studied in great detail the properties of Chinese porcelain and the physics of iron and steel.

[19] Unlike the technologies that developed in Europe and the United States in the second half of the nineteenth century, science, in this view, had little direct guidance to offer to the Industrial Revolution (Hall, 1974, p. 151). Shapin notes that "it appears unlikely that the 'high theory' of the Scientific Revolution had any substantial direct effect on economically useful technology either in the seventeenth century or in the eighteenth.... historians have had great difficulty in establishing that any of these spheres of technologically or economically inspired science bore

connection becomes gradually tighter, yet the influence of science proper on some branches of production (and by no means all at that) does not become decisive until after 1870.

All the same, the success of the Industrial Revolution must be found in the developments in the area of useful knowledge that occurred in Europe before and around 1750. What mattered was not so much scientific knowledge itself but the method and culture involving the generation and diffusion of propositional knowledge. The Industrial Revolution and its aftermath were based on a set of propositional knowledge that was not only increasing in size, but which was becoming increasingly accessible, and in which segments that were more effective were becoming tighter. Propositional knowledge was increasingly tested by whether the techniques that were based on it could be verified, either by experiment or by virtue of the performance of the techniques based on them.

The Scientific Method that evolved in the seventeenth century meant that observation and experience were placed in the public domain. Betty Jo Dobbs (1990), William Eamon (1990, 1994), and more recently Paul David (1997) have pointed to the scientific revolution of the seventeenth century as the period in which "open science" emerged, when knowledge about the natural world became increasingly nonproprietary and scientific advances and discoveries were freely shared with the public at large. Thus scientific knowledge became a public good, communicated freely rather than confined to a secretive exclusive few as had been the custom in medieval Europe. The sharing of knowledge within "open science" required systematic reporting of methods and materials using a common vocabulary and consensus standards, and should be regarded as an exogenous decline in access costs, which made the propositional knowledge, such as it was, available to those who might find a use for it. Those who added to useful knowledge would be rewarded by honor, peer recognition, and fame – not a monetary reward that was in any fashion proportional to their contribution. Even those who discovered matters of significant insight to industry, such as Claude Berthollet, Joseph Priestly, and Humphry Davy, wanted credit, not profit.

Scientific Method here also should be taken to include the changes in the rhetorical conventions that emerged in the seventeenth century,

---

substantial fruits" (1996, pp. 140–41, emphasis added).

during which persuasive weight continued to shift away from "authority" toward empirics, but which also increasingly set the rules by which empirical knowledge was to be tested so that useful knowledge could be both accessible and trusted.[20] Verification meant that a deliberate effort was made to make useful knowledge tighter and thus more likely to be used. It meant a willingness, rarely observed before, to discard old and venerable interpretations and theories when they could be shown to be in conflict with the evidence. Scientific method meant that a class of experts evolved who often would decide which technique worked best.[21]

The other crucial transformation that the Industrial Revolution inherited from the seventeenth century was the growing change in the very purpose and objective of propositional knowledge. Rather than proving some religious point, such as illustrating the wisdom of the creator, or the satisfaction of that most creative of human characteristics, curiosity, natural philosophers in the eighteenth century increasingly came under the influence of the idea that the main purpose of knowledge was to improve mankind's material condition – that is, find technological applications. Bacon in 1620 had famously defined technology by declaring that the control of humans over things depended on the accumulated knowledge about how nature works, since "she was only to be commanded by obeying her." This idea was of course not entirely new, and traces of it can be found in medieval thought and even in Plato's *Timaeus*, which proposed a rationalist view of the universe and was widely read by twelfth-century intellectuals. In the seventeenth century, however, the practice of science

---

[20] Shapin (1994) has outlined the changes in trust and expertise in Britain during the seventeenth century associating expertise, for better or for worse, with social class and locality. While the approach to science was ostensibly based on a "question authority" principle (the Royal Society's motto was *nullius in verba*—on no one's word), in fact no system of useful (or any kind of) knowledge can exist without some mechanism that generates trust. The apparent skepticism with which scientists treated the knowledge created by others increased the trust that others had in the findings, because outsiders could then assume —as is still true today—that these findings had been scrutinized and checked by other "experts."

[21] As Hilaire-Pérez (2000, p. 60) put it, "the value of inventions was too important an economic stake to be left to be dissipated among the many forms of recognition and amateurs: the establishment of truth became the professional responsibility of academic science."

became increasingly permeated by the Baconian motive of material progress and constant improvement, attained by the accumulation of knowledge.[22] The founding members of the Royal Society justified their activities by their putative usefulness to the realm. There was a self-serving element to this, of course, much as with National Science Foundation grant proposals today. Practical objectives in the seventeenth century were rarely the primary objective of the growth of formal science. But part of the changing culture implied a gradual change in the agenda of research.

A central intellectual change in Europe before the Industrial Revolution, oddly neglected by economic historians, was the Enlightenment. Definitions of this amorphous and often contradictory historical phenomenon are many, but for the purposes of explaining the Industrial Revolution we only to examine a slice of it, which I have termed the *Industrial Enlightenment*. To be sure, some historians have noted the importance of the Enlightenment as a culture of rationality, progress, and growth through knowledge.[23] Perhaps the most widely diffused Enlightenment view involved the notion that long-term social improvement was possible. It surely is true that not all Enlightenment philosophers believed that progress was either desirable or inevitable. And yet their work created the attitudes, the institutions, and the mechanisms by which new useful knowledge was created, spread, and put to good use. Above all was the pervasive belief in the Baconian notion that we can attain material progress (that is, economic growth) through controlling nature and that we can only

---

[22] Robert K. Merton ([1938] 1970, pp. ix, 87) asked rhetorically how "a cultural emphasis upon social utility as a prime, let alone an exclusive criterion for scientific work affects the rate and direction of advance in science" and noted that "science was to be fostered and nurtured as leading to the improvement of man's lot by facilitating technological invention." He might have added that non-epistemic goals for useful knowledge and science, that is to say, goals that transcend knowledge for its own sake and look for some application, affected not only the rate of growth of the knowledge set but even more the chances that existing knowledge will be translated into techniques that actually increase economic capabilities and welfare.

[23] One of the most cogent statements is in McNeil (1987, pp. 24-25) who notes the importance of a "faith in science that brought the legacy of the Scientific Revolution to bear on industrial society ... it is imperative to look at the interaction between culture *and* industry, between the Enlightenment and the Industrial Revolution."

harness nature by understanding her. Francis Bacon, indeed, is a pivotal figure in understanding the Industrial Enlightenment and its impact. Modern scholars seem agreed: Bacon was the first to regard knowledge as subject to constant growth, an entity that continuously expands and adds to itself rather than concerned with retrieving, preserving and interpreting old knowledge (Vickers, 1992, esp. pp. 496-97).[24] The understanding of nature was a social project in which the division of knowledge was similar to Adam Smith's idea of the division of labor, another enlightenment notion.[25] Bacon's idea of bringing this about was through what he called a "House of Salomon" – a research academy in which teams of specialists collect data and experiment, and a higher level of scientists tty to distill these into general regularities and laws. Such an institution was the Royal Society, whose initial objectives were inspired by Lord Bacon.[26]

Nothing of the sort, I submit, can be detected in the Ottoman Empire, India, Africa, or China. It touched only ever so lightly (and with a substantial delay) upon Iberia, Russia, and South America but in many of these areas it encountered powerful resistance and retreated. Invention, as many scholars have rightly stressed, was never a European monopoly, and much of its technological creativity started with adopting ideas and tech-

---

[24] Bacon was pivotal in inspiring the Industrial Enlightenment. His influence on the Industrial Enlightenment can be readily ascertained by the deep admiration the encyclopédistes felt toward him, including a long article on Baconisme written by the Abbé Pestre and the credit given him by Diderot himself in his entries on *Art* and *Encyclopédie*. The *Journal Encyclopédique* wrote in 1756 "If this society owes everything to Chancellor Bacon, the philosopher doe not owe less to the authors of the *Encyclopédie*" (cited by Kronick, 1962, p. 42). The great Scottish Enlightenment philosophers Dugald Stewart and Francis Jeffrey agreed on Baconian method and goals, even if they differed on some of the interpretation (Chitnis, 1976, pp. 214-15).

[25] A typical passage in this spirit was written by the British chemist and philosopher Joseph Priestley (1768, p. 7): "If, by this means, one art or science should grow too large for an easy comprehension in a moderate space of time, a commodious subdivision will be made. Thus all knowledge will be subdivided and exteded, and *knowledge* as Lord Bacon observes, being *power*, the human powers will be increased ... men will make their situation in this world abundantly more easy and comfortable."

[26] McClellan (1985), p. 52. It should be added that strictu sensu the Royal Society soon allowed in amateurs and dilettantes and thus became less of a pure "Baconian" institution than the French Académie Royale.

niques the Europeans had observed from others (Mokyr, 1990). The difference was the ability to break out of the circle of concavity and negative feedback and smash the upper bound on income that the limitations of knowledge and institutions had set on practically all economies until then. The stationary state was replaced by the steady state. It is this phenomenon rather than coal or the ghost acreage of colonies that answers Pomeranz's query (2000, p. 48) why Chinese science and technology – which did not "stagnate" – "did not revolutionize the Chinese economy."

The Industrial Enlightenment can be viewed in part as a movement that insisted on asking not just "which techniques work" but also "why" – realizing that such questions held the key to continuing progress. In that sense, the intellectuals at its center felt intuitively that constructing and widening an epistemic base for the techniques in use would lead to continuing technological progress. Scientists, engineers, chemists, medical doctors, and agricultural improvers made sincere efforts to generalize from the observations they made, to connect observed facts and regularities (including successful techniques) to the formal propositional knowledge of the time, and thus provide the techniques with wider epistemic bases. The bewildering complexity and diversity of the world of techniques in use was to be reduced to a finite set of general principles governing them.[27] These insights would lead to extensions, refinements, and improvements, as well as speed up and streamline the process of invention.[28] Asking such questions was of course much easier than answering them. In the longer term, however, asking such questions and developing the tools to get to the answers were essential if technical progress was not to fizzle out.[29] The

---

[27] Thus Erasmus Darwin, grandfather of the biologist and himself a charter member of the Lunar Society and an archtypical member of the British Industrial Enlightenment, complained in 1800 that Agriculture and Gardening had remained only Arts without a true theory to connect them (Porter, 2000, p. 428). For details about Darwin, see especially McNeil (1987) and Uglow (2002).

[28] Somewhat similar views have been expressed recently by other scholars such as John Graham Smith (2001) and Picon (2001).

[29] George Campbell, an important representative of the Scottish Enlightenment noted that "All art [including mechanical art or technology] is founded in science and practical skills lack complete beauty and utility when they do not originate in knowledge" (cited by Spadafora, 1990, p. 31).

typical enlightenment inventor did more than tinkering and trial-and-error fiddling with existing techniques, he tried to relate puzzles and challenges to whatever general principles could be found, and if necessary to formulate such principles anew. To do so, each inventor needed some mode of communication that would allow him to tap the knowledge of others. The paradigmatic example of such an inventor remains the great James Watt, whose knowledge of mathematics and physics were matched by his tight connections to the best scientific minds of his time, above all Joseph Black and Joseph Priestley. The list of slightly less famous pioneers of technology can be made arbitrarily long.

The other side of the Industrial Enlightenment had to do with the diffusion of and the access to existing knowledge. The *philosophes* fully realized that knowledge should not be confined to a select few but should be disseminated as widely as possible. Some Enlightenment thinkers believed this was already happening: the philosopher and psychologist David Hartley believed that "the diffusion of knowledge to all ranks and orders of men, to all nations, kindred and tongues and peoples... cannot be stopped but proceeds with an ever accelerating velocity."[30] Diffusion needed help, however, and much of the Industrial Enlightenment was dedicated to making access to useful knowledge easier and cheaper.[31] From the widely felt need to rationalize and standardize weights and measures, the insistence on writing in vernacular language, to the launching of scientific societies and academies (functioning as de facto clearing houses of useful knowledge), to that most paradigmatic Enlightenment triumph, the *Grande Encyclopédie*, the notion of diffusion found itself at the center of attention among intellectuals.[32] Precisely because the Industrial Enlightenment was

---

[30] Cited by Porter (2000, p. 426).

[31] The best summary of this aspect of the Industrial Enlightenment was given by Diderot in his widely-quoted article on "Arts" in the *Encyclopédie*: "We need a man to rise in the academies and go down to the workshops and gather material about the [mechanical] arts to be set out in a book that will persuade the artisans to read, philosophers to think along useful lines, and the great to make at least some worthwhile use of their authority and wealth."

[32] Roche (1998, pp. 574-75) notes that "if the *Encyclopédie* was able to reach nearly all of society (although ... peasants and most of the urban poor had access to the work only indirectly), it was because the project was broadly conceived as a work of

not a national or local phenomenon, it became increasingly felt that differences in language and standards became an impediment and increased access costs. Watt, James Keir, the Derby clockmaker John Whitehurst, all worked on a system of universal terms and standards, that would make French and British experiments "speak the same language" (Uglow, 2002, p. 357). Books on science and technology were translated rather quickly, even when ostensibly Britain and France were at war with one another.

Access costs depended in great measure on knowing what was known, and for that search engines were needed. The ultimate search engine of the eighteenth century was the encyclopedia. Diderot and d'Alembert's *Encyclopédie* did not augur the Industrial Revolution, it did not predict factories, and had nothing to say about mechanical cotton spinning equipment or steam engines. It catered primarily to the landowning elite and the bourgeoisie of the *ancien régime* (notaries, lawyers, local officials) rather than specifically to an innovative industrial bourgeoisie, such as it was. It was, in many ways, a conservative document (Darnton, 1979, p. 286). But the Industrial Enlightenment, as embodied in the *Encyclopédie* and similar works that were published in the eighteenth century did propose a very different way of looking at technological knowledge: instead of intuition came systematic analysis, instead of mere dexterity an attempt to attain an understanding of the principles at work, instead of secrets learned from a master, an open and accessible system of training and learning. It was also a comparatively user-friendly compilation, arranged in an accessible way, and while its subscribers may not have been mostly artisans and small manufacturers, the knowledge contained in it dripped down through a variety of leaks to those who could make use of it.[33]

---

popularization, of useful diffusion of knowledge." The cheaper versions of the Diderot-d'Alembert masterpiece, printed in Switzerland, sold extremely well: the Geneva (quarto) editions sold around 8000 copies and the Lausanne (octavo) editions as many as 6000.

[33] Pannabecker points out that the plates in the *Encyclopédie* were designed by the highly skilled Louis-Jacques Goussier who eventually became a machine designer at the Conservatoire des arts et métiers in Paris (Pannabecker, 1996). They were meant to popularize the rational systematization of the mechanical arts to facilitate technological progress.

Encyclopedias and "dictionaries" were supplemented by a variety of textbooks, manuals, and compilations of techniques and devices that were somewhere in use. The biggest one was probably the massive *Descriptions des arts et métiers* produced by the French Académie Royale des Sciences.[34] Specialist compilations of technical and engineering data appeared, such as the detailed descriptions of windmills (*Groot Volkomen Moolenboek*) published in the Netherlands as early as 1734. A copy was purchased by Thomas Jefferson (Davids, 2001). Jacques-François Demachy's *l'Art du distillateur d'eaux fortes* (1773) (published as a volume in the *Descriptions*) is a "recipe book full of detailed descriptions of the construction of furnaces and the conduct of distillation" (John Graham Smith, 2001, p. 6). In agriculture, meticulously compiled data collections looking at such topics as yields, crops, and cultivation methods were common.[35]

The Industrial Enlightenment realized instinctively that one of the great sources of technological stagnation was a social divide between those who knew things ("*savants*") and those who made things ("*fabricants*"). To construct pipelines through which those two groups could communicate was at the very heart of the movement.[36] The relationship between those who possessed useful knowledge and those who might find a use for it was

---

[34] The set included 13,500 pages of text and over 1,800 plates describing virtually every handicraft practiced in France at the time, and every effort was made to render the descriptions "realistic and practical" (Cole and Watts, 1952, p. 3).

[35] William Ellis's *Modern Husbandman or Practice of Farming* published in 1731 gave a mont-by-month set of suggestions, much like Arthur Young's most successful book, *The Farmer's Kalendar* (1770). Most of these writings were empirical or instructional in nature, but a few actually tried to provide the readers with some systematic analysis of the principles at work. One of those was Francis Home's *Principles of Agriculture and Vegetation* (1757). One of the great private data collection projects of the time were Arthur Young's famed *Tours* of various parts of England and William Marshall's series on *Rural Economy* (Goddard, 1989). They collected hundreds of observations on farm practice in Britain and the continent. although at times Young's conclusions were contrary to what his own data indicated (see Allen and Ó Gráda, 1988).

[36] This point was first made by Zilsel(1942) who placed the beginning of this movement in the middle of the sixteenth century. While this may be too early for the movement to have much economic effect, the insight that technological progress occurs when intellectuals communicate with producers is central to its historical explanation.

changing in eighteenth-century Europe and points to a reduction in access costs. They also served as a mechanism through which practical people with specific technical problems to solve could air their needs and thus influence the agenda of the scientists, while at the same time absorbing what best-practice knowledge had to offer. The movement of knowledge was thus bi-directional, as perhaps seems natural to us in the twenty-first century. In eighteenth-century Europe, however, such exchanges were still quite novel.

An interesting illustration can be found in the chemical industry. Pre-Lavoisier chemistry, despite its limitations, is an excellent example of how *some* knowledge, no matter how partial or erroneous, was believed to be of use in mapping into new techniques. The pre-eminent figure in this field was probably William Cullen, a Scottish physician and chemist. Cullen lectured (in English) to his medical students, but many outsiders connected with the chemical industry audited his lectures. Cullen believed that as a philosophical chemist he had the knowledge needed to rationalize the processes of production (Donovan, 1975, p. 78). He argued that pharmacy, agriculture, and metallurgy were all "illuminated by the principles of philosophical chemistry" and added that "wherever any art [that is, technology] requires a matter endued with any peculiar physical properties, it is chemical philosophy which informs us of the natural bodies possessed of these bodies" (cited by Brock, 1992, pp. 272–73).[37] He and his colleagues worked, among others, on the problem of purifying salt (needed for the Scottish fish-preservation industry), and that of bleaching with lime, a common if problematic technique in the days before chlorine. This kind of work "exemplifies all the virtues that eighteenth-century chemists believed would flow from the marriage of philosophy and practice" (Donovan, 1975, p. 84).

Ironically, this marriage remained barren for many decades. In chemistry the expansion of the epistemic base and the flurry of new techniques it generated did not occur fully until the mid-nineteenth century (Fox, 1998). Cullen's prediction that chemical theory would yield the principles that would direct innovations in the practical arts remained, in the words of the leading expert on eighteenth-century chemistry, "more

---

[37] Very similar sentiments were expressed by the author of the article on chemistry, Gabriel-François Venel, in the *Encyclopédie.* He regarded advances in arts and chemical science as reciprocal, bound together on a common trunk (Keyser, 1990, p. 228).

in the nature of a promissory note than a cashed-in achievement" (Golinski, 1992, p. 29). Manufacturers needed to know why colors faded, why certain fabrics took dyes more readily than others, and so on, but as late as 1790 best-practice chemistry was incapable of helping them much (Keyser, 1990, p. 222). Before the Lavoisier revolution in chemistry, it just could not be done, no matter how suitable the social climate: the epistemic base simply did not exist. All the same, Cullen personifies a social movement that increasingly sought to increase its propositional knowledge for economic purposes, a personification of scientific culture. Whether or not he could deliver, his patrons and audience in the culture of the Scottish Enlightenment believed that there was a chance he could (Golinski, 1988) and put their money behind their beliefs.

To dwell on one more example, consider the development of steam power. The ambiguities of the relations between James Watt and his mentor, the Scottish scientist Joseph Black are well-known. Whether or not Watt's crucial insight of the separate condenser was due to Black's theory of latent heat, there can be little doubt that the give-and-take between the scientific community in Glasgow and the creativity of men like Watt was essential in smoothing the path of technological progress. Richard Trevithick, the Cornish inventor of the high pressure engine, posed sharp questions to his scientist acquaintance Davies Gilbert (later President of the Royal Society), and received answers that supported and encouraged his work (Burton, 2000, pp. 59-60). Decades later, the work of Joule and Rankine on thermodynamics led to the development of the two cylinder compound marine steam engine.[38]

As might be expected, in some cases the bridge between propositional and prescriptive knowledge occurred within the same mind: the very same people who also were contributing to science made some critical inventions (even if the exact connection between their science and their ingenuity is not always clear). The importance of such dual or "hybrid" careers, as Eda Kranakis (1992) has termed them, is that access to the propositional knowledge that could underlie an invention is immediate, as is the feedback to propositional knowledge. In most cases the technology

---

[38] In other areas, too, the discourse between those who controlled S-knowledge and those who built new techniques was fruitful. Henry Cort, whose invention of the puddling and rolling process was no less central than Watt's separate consenser, also consulted Joseph Black during his work.

shaped the propositional research as much as the other way around. The idea that those contributing to propositional knowledge should specialize in research and leave its "mapping" into technology to others had not yet ripened. Among the inventions made by people whose main fame rests on their scientific accomplishments were the chlorine bleaching process invented by the chemist Claude Berthollet, the invention of carbonated (sparkling) water and rubber erasers by Joseph Priestley, and the mining safety lamp invented by the leading scientist of his age, Humphry Davy (who also, incidentally, wrote a textbook on agricultural chemistry and discovered that a tropical plant named *catechu* was a useful additive to tanning).[39]

Typical of the "dual career" phenomenon was Benjamin Thompson (later Count Rumford, 1753-1814), an American-born mechanical genius who was on the loyalist side during the War of Independence and later lived in exile in Bavaria, London, and Paris; he is most famous for the scientific proof that heat is not a liquid (known at the time as *caloric*) that flows in and out of substances. Yet Rumford was deeply interested in technology, helped establish the first steam engines in Bavaria, and invented (among other things) the drip percolator coffeemaker, a smokeless-chimney Rumford stove, and an improved oil lamp. He developed a photometer designed to measure light intensity and wrote about science's ability to improve cooking and nutrition (G. I. Brown, 1999, pp. 95–110). Rumford is as good a personification of the Industrial Enlightenment as one can find. Indifferent to national identity and culture, Rumford was a "Westerner" whose world spanned the entire northern Atlantic area (despite being an exile from the United States, he left much of his estate to establish a professorship at Harvard). In that respect he resembled his older compatriot inventor Benjamin Franklin, who was as celebrated in Britain and France as he was in his native Philadelphia. Rumford could, within the same mind, map from his knowledge of natural phenomena and regularities to create

[39] It is unclear how much of the best-practice science was required for the safety lamp, and how much was already implied by the empirical propositional knowledge accumulated in the decades before 1815. It is significant that George Stephenson, of railway fame, designed a similar device at about the same time.

things he deemed useful for mankind (Sparrow, 1964, p. 162).[40] Like Franklin and Davy, he refused to take out a patent on any of his inventions — as a true child of the Enlightenment he was committed to the concept of open and free knowledge.[41] Instead, he felt that honor and prestige were often a sufficient incentive for people to contribute to useful knowledge. He established the Rumford medal, to be awarded by the Royal Society "in recognition of an outstandingly important recent discovery in the field of thermal or optical properties of matter made by a scientist working in Europe, noting that Rumford was concerned to see recognised discoveries that tended to promote the good of mankind." Not all scientists eschewed such profits: the brilliant Scottish noblemen Archibald Cochrane (Earl of Dundonald) made a huge effort to render the coal tar process he patented profitable, but failed and ended up losing his fortune.

The other institutional mechanism emerging during the Industrial Enlightenment to connect between those who possessed prescriptive knowledge and those who wanted to apply it was the emergence of meeting places where men of industry interacted with natural philosophers. So-called scientific societies, often known confusingly as literary and philosophical societies, sprung up everywhere in Europe. They organized lectures, symposia, public experiments, and discussion groups, in which the topics of choice were the best pumps to drain mines, or the advantages of growing clover and grass.[42] Most of them published some form of "proceedings," as often meant to popularize and diffuse existing knowledge as it was to dispay new discoveries. Before 1780 most of these societies were

---

[40] It is telling that Rumford helped found the London Royal Institute in 1799. This institute was explicitly aimed at the diffusion of useful knowledge to wider audiences through lectures. In it the great Humphry Davy and his illustrious pupil Michael Faraday gave public lectures and did their research.

[41] The most extreme case of a scientist insisting on open and free access to the propositional knowledge he discovered was Claude Berthollet, who readily shared his knowledge with James Watt, and declined an offer by Watt to secure a patent in Britain for the exploitation of the bleaching process (J. G. Smith, 1979, p. 119).

[42] The most famous of these societies were the Manchester Literary and Philosophical Society (founded in 1781) and the Birmingham Lunar Society, where some of the great entrepreneurs and engineers of the time mingled with leading chemists, physicists, and medical doctors. But in many provincial cities such as Liverpool, Hull, and Bradford, a great deal of activity took place.

informal and ad hoc, but they eventually became more formal. The British Society of Arts, founded in 1754, was a classic example of an organization that embodied many of the ideals of the Industrial Enlightenment. Its purpose was "to embolden enterprise, to enlarge science, to refine art, to improve manufacture and to extend our commerce." Its activities included an active program of awards and prizes for successful inventors: over 6,200 prizes were granted between 1754 and 1784.[43] The society took the view that patents were a monopoly, and that no one should be excluded from useful knowledge. It therefore ruled out (until 1845) all persons who had taken out a patent from being considered for a prize and even toyed with the idea of requiring every prize-winner to commit to never take out a patent.[44] It served as a clearing house for technological information, reflecting the feverish growth of supply and demand for useful knowledge.

What was true for Britain was equally true for Continental countries affected by the Enlightenment. In the Netherlands, rich but increasingly technologically backward, heroic efforts were established to infuse the economy with more innovativeness.[45] In Germany, provincial academies were founded in all the significant German states in the eighteenth century. The Berlin Academy, in its early years directed by the great Leibniz, and among its achievements wa sthe discovery that sugar could be extracted from beets. In France, great institutions were created

---

[43] For details see, Wood (1913), Hudson and Luckhurst (1954).

[44] Hilaire-Pérez (2000), p. 197. Wood (1913), pp. 243-45.

[45] The first of these was established in Haarlem in 1752, and within a few decades the phenomenon spread much like in England to the provincial towns. The Scientific Society of Rotterdam known oddly as the *Batavic Association for Experimental Philosophy* was the most applied of all, and advocated the use of steam engines (which were purchased in the 1770s but without success). The Amsterdam Society was known as *Felix Meritis* and carried out experiments in physics and chemistry. These societies stimulated interest in physical and experimental sciences in the Netherlands, and they organized prize-essay contests on useful applications of natural philosophy. A physicist named Benjamin Bosma for decades gave lectures on mathematics, geography, and applied physics in Amsterdam. A Dutch Society of Chemistry founded in the early 1790s helped to convert the Dutch to the new chemistry proposed by Lavoisier (Snelders, 1992). The Dutch high schools, known as *Athenea* taught mathematics, physics, astronomy, and at times counted distinguished scientists among their staff.

under royal patronage, above all the Académie Royale des Sciences, created by Colbert and Louis XIV in 1666 to disseminate information and resources.[46] Yet the phenomenon was nationwide: 33 official learned societies were functioning in the French provinces during the eighteenth century counting over 6,400 members. Overall, McClellan (1981, p. 547) estimates that during the century perhaps between 10,000 and 12,000 men belonged to learned societies that dealt at least in part with science. Between 1692 and 1792, 11 towns in Italy formed scientific societies (Inkster, 1991, p. 35). At the level of the creation of propositional knowledge, at least, there is little evidence that the *ancien régime* was incapable of generating sustained progress.

The *Académie Royale* exercised a fair amount of control over the direction of French scientific development and acted as technical advisor to the monarchy. By determining what was published and exercising control over patents, the *Académie* became a powerful administrative body, providing scientific and technical advice to government bureaus. France, of course, had a somewhat different objective than Britain: it is often argued that the *Académie* linked the aspirations of the scientific community to the utilitarian concerns of the government creating not a Baconian society open to all comers and all disciplines but a closed academy limited primarily to Parisian scholars (McClellan, 1981). Yet the difference between France and Britain was one of emphasis and nuance, not of essence: they shared a utilitarian optimism of man's ability to create wealth. In Germany, too, learned academies and private societies were founded to promote industrial, agricultural, and political progress through science. Around 200 societies appeared during the half-century spanning from the Seven Years War to the climax of the Napoleonic occupation of Germany, such as the Patriotic Society founded at Hamburg in 1765 (Lowood, 1991, pp. 26-27). These societies, too, emphasized the welfare of the population at large and the

---

[46] It was one of the oldest and financially best supported scientific societies of the eighteenth century, with a membership which included d'Alembert, Buffon, Clairaut, Condorcet, Fontenelle, Laplace, Lavoisier, and Reaumur. It published the most prestigious and substantive scientific series of the century in its annual proceedings Histoire et Memoires and sponsored scientific prize contests such as the Meslay prizes. It recognized achievement and rewarded success for individual discoveries and enhanced the social status of scientists, granting salaries and pensions. A broad range of scientific disciplines were covered, with mathematics and astronomy particularly well represented, as well as botany and medicine.

country over private profit. Local societies supplemented and expanded the work of learned national academies, which were first founded in Berlin in 1700.[47] Publishing played an important role in the work of societies bent on the encouragement of invention, innovation and improvement. This reflected the emergence of open knowledge, a recognition that knowledge was a non-rivalrous good the diffusion of which was constrained by access costs.

To summarize, then, the Industrial Revolution had intellectual roots that needed to be met if sustained economic growth could take place just as it had to satisfy economic and social conditions. The importance of property rights, incentives, factor markets, natural resources, law and order, market integration, and many other economic elements is not in question. But we need to realize that without understanding the growth of useful knowledge, the technological elements will remain inside a black box.

The emergence of technological modernity.

The essence of technological modernity is non-stationarity: many scholars have observed that technological change has become self-propelled and autocatalytic, in which change feeds on change. Unlike other forms of growth, spiraling technological progress does not appear to be bounded from above. Predictions in the vein of "everything that can be invented already has been" have been falsified time and again. The period that followed the Industrial Revolution was one in which innovation intensified, and while we can distinguish a certain ebb and flow, in which major breakthroughs and a cluster of macroinventions were followed by waves of

---

[47] The German local societies were private institutions, unlike state-controlled academies, which enabled them to be more open, with few conditions of entry, unlike the selective, elitist academies. They broke down social barriers, for the established structures of Old Regime society might impede useful work requiring a mixed contribution from the membership of practical experience, scientific knowledge, and political power. Unlike the more scientifically-inclined academies, they invited anyone to join, such as farmers, peasants, artisans, craftsmen, foresters, and gardeners, and attempted to improve the productivity of these occupations and solve the economic problems of all classes. Prizes rewarded tangible accomplishments, primarily in the agricultural or technical spheres. Their goal was not to advance learning like earlier academies, but to apply useful results of human knowledge, discovery and invention to practical and civic life (Lowood, 1991).

microinventions and secondary extensions and applications, it is clear that the modern era is one in which rapid and perhaps accelerating change has become the norm. In the premodern past, whether in Europe or elsewhere in the world, invention had remained the exception, if perhaps not an uncommon one. In the second half of the nineteenth century and even more so in the twentieth century, change has become the norm, and even in areas previously untouched by technological innovation, mechanization, automation, and novelty have become inevitable. There is no evidence to date that technology in its widest sense converges to anything.

To oversimplify, the Industrial Revolution could be reinterpreted in light of the changes in the characteristics and structure of propositional knowledge in the eighteenth century and the techniques that rested on it. Before 1750 the human race, as a collective, did not know enough to generate the kind of sustained technological progress that could account for the growth rates we observe. In the absence of such knowledge, no set of institutions, no matter how benevolent, could have substituted for useful knowledge. The dynamics of knowledge itself were critical to the historical process. Useful knowledge after the Industrial Revolution increased by feeding on itself, spinning out of control as it were, whereas before it had always been limited by its epistemic base and suppressed by economic and social factors, creating what physicists call a "phase transition."[48]

How do we explain this change in technological dynamic? In economics, phase transitions can be said to occur when a dynamic system has multiple steady states such as an economy that has a "poverty trap" (low-income equilibrium) and a high income (or rapid growth steady state). A phase transition occurs when the system switches from one equilibrium or regime to another. A simple model in which this can be illustrated is one in which capital and skills are highly complementary. In such models one equilibrium is characterized by rapid investment, which raises the demand for skills; the positive feedback occurs because the increase in the rate of return to human capital induces parents to invest more in their children, have fewer children (since they become more expensive), which raises the rate of return on physical capital even more and encourages investment. A regime change may occur when an exogenous shock is violent enough to

---

[48] For a definition of phase transitions, see for instance Ruelle (1991), pp. 122-23.

bump the system off one basin of attraction and move it to another one. The difficulty is to identify a historical shock that was sufficiently powerful to "bump" the system to a rapid growth trajectory. The alternative is to presume that historical processes cause the underlying parameters to change slowly but cumulatively, until one day what was a slow-growth steady state is no longer an equilibrium at all and the system, without a discernible shock, moves rather suddenly into a very different steady state. These models, pioneered by Galor and Weil (2000), move from comparative statics with respect to a parameter determining the dynamic structure, to a dynamical system in which this parameter is a latent state variable that evolves and ultimately can generate a phase transition.[49]

Recent work in growth theory have produced a class of models that reproduce this feature in one form or another. Three models of interest have this characteristic, even if they include additional elements. Cervellati and Sunde (2002) model the relationship between mortality and human capital investment. This is a little explored aspect of modernization, but one that must have been of some importance. All other things equal, longer life expectancy would encourage investment in human capital, although it is important to emphasize that a reduction in infant mortality would not directly bring this about, because decisions about human capital are made later in life. Increases in life expectancy at age 10 or so are more relevant here. The other important idea in the Cervellati and Sunde paper is that human capital comes in two forms, a "theoretical" form and a "practical" form, corresponding roughly to "scientific" and "artisanal" knowledge or the categories of useful knowledge proposed above. They assume that human abilities are heterogeneous but that the threshold a at which people start to invest in "theoretical" knowledge as opposed to "crafts" is endogenously determined by life expectancy. This threshold level depends on the costs of acquiring the two types of human capital, their rates of return, and the life expectancy over which they are amortized. Given their

[49] Another example of this type of "phase transition" has been proposed recently by David (1998). He envisages the community of "scientists" to consist of local networks or "invisible colleges" in the business of communicating with each other. Such transmission between connected units can be modeled using percolation models in which information is diffused through a network with a certain level of connectivity. David notes that these models imply that there is a minimum level of persistently communicative behavior that a network must maintain for knowledge to diffuse through and that once this level is achieved the system becomes self-sustaining.

assumptions, the locus of points in the life-expectancy-ability space that define an intra-generational equilibrium is S-shaped. A second relationship in this model is that life expectancy itself depends on the level of education of the previous generation: better educated parents will be better situated to help their children survive. The model is closed by postulating a relationship between the high quality human capital and total productivity. Under these conditions, the dynamic evolution of human capital formation and life expectancy *for any level of technology* can be shown in the following diagram (see fig. 1, adapted from Cervellati and Sunde). This model reproduces a sophisticated version of a poverty "trap" at the attractor point Z (where life expectancy T is low and the threshold level of *a* is close to unity, so that few people choose the high-quality education) and the highly-developed equilibrium point Y in which life expectancy is higher and more people get the high quality education. The neat aspect of the Cervellati-Sunde model is that if for some reason the productivity of the high-quality human capital rises, the curve HH rotates counter-clockwise and becomes less S-shaped. This produces the kind of observed phase transition when the old poverty trap is no longer an equilibrium and the system abruptly starts to move to point Y [see fig. 2].

In their model, the shift is built-in, because if only a few people are receiving the "theoretical" education, they will expand productivity. But any exogenous disturbance that raises the marginal productivity of "scientific activity" will have the same effect, including an exogenous increase in the stock of propositional knowledge and an ideologically-induced change in the agenda of research. Clearly, then, the Industrial Enlightenment, much like an endogenous growth in productivity, can produce a sequence as in fig. 2 and generate an "Industrial Revolution" of this type. While under the assumptions of this paper an Industrial Revolution is "inevitable," they recognize that if technological progress has stochastic elements, this could imply a different prediction (p. 23). Either way, however, the emergence of technologically-based "modern growth" can be understood without the need for a sudden violent shock.

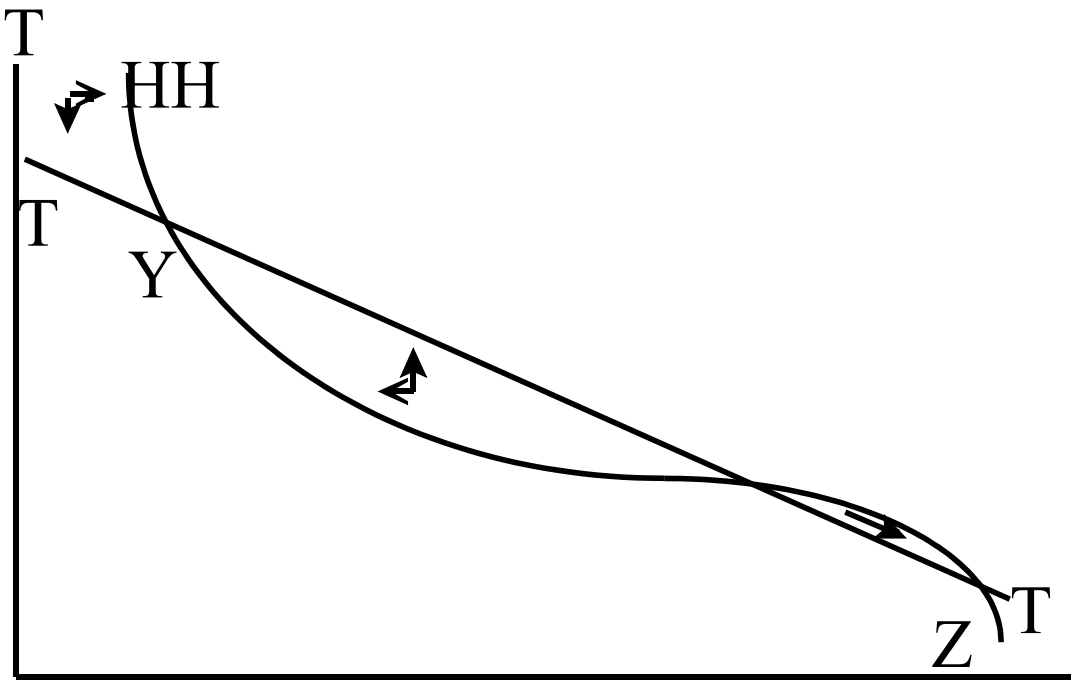


Fig. 1: Poverty trap in a traditional society society a
Source: Cervellati and Sunde (2003)

A second model that generates a sudden transition of this nature is that by Galor and Weil (2000). The Galor-Weil model purports to explain both the Industrial Revolution (post-1750 technological progress) *and* the demographic transition (post-1850 fertility decline). The basic idea is that technological progress itself (rather than the higher income it creates) raises the rate of return on human capital. One interesting idea in this paper is that technological progress is actually costly because it devalues the skills that parents can impart on their children, and instead needs a more institutionalized form of education in which new technology can be "taught." In this they follow an important idea originally proposed by T.W. Schultz (1975), that one of the main functions of human capital is to help

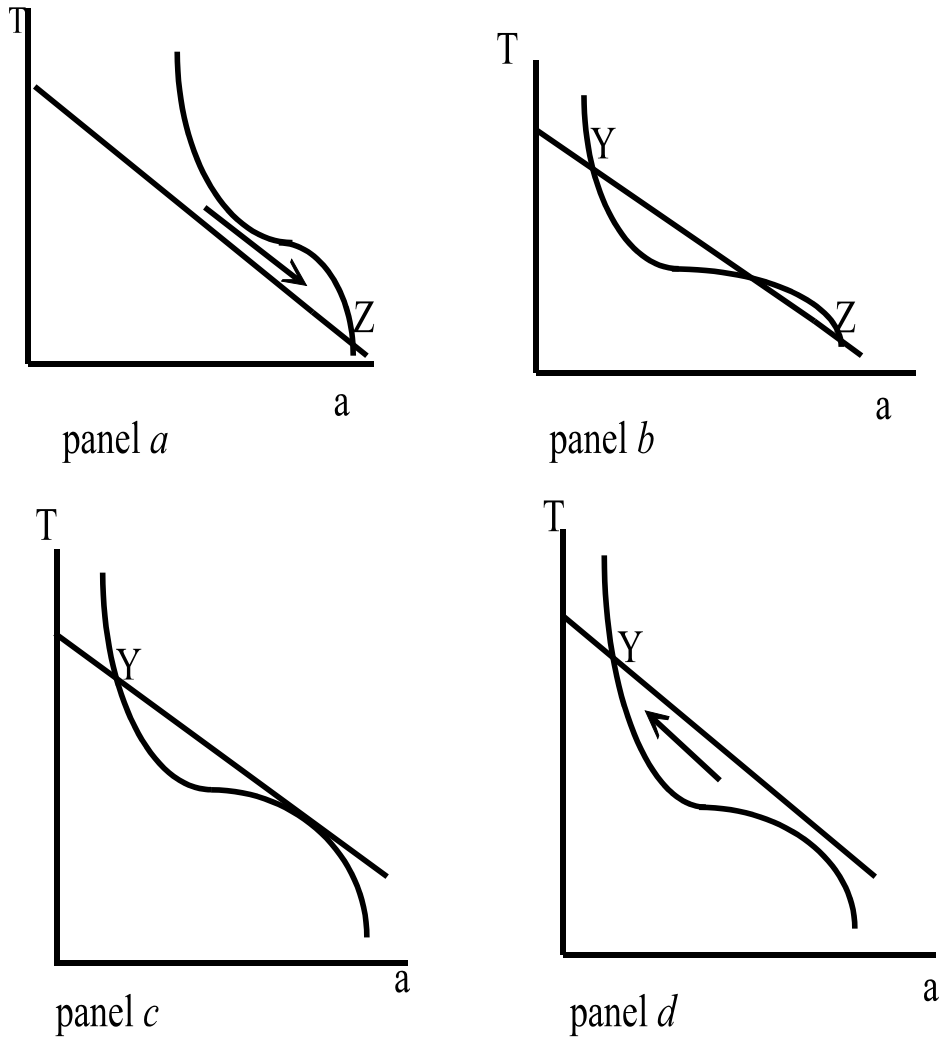


Fig. 2: Transition from traditional to modern society

Source: Cervellati and Sunde (2003)

individuals cope with disequilibria. The positive feedback effect occurs because investment in human capital generates more technological progress. In the Galor-Weil model, the economic *ancien régime* is not really a steady state but a "pseudo steady state" despite its long history: within a seeming stability the seeds for the phase transition are germinating invisibly. The way this works is that technological progress in the pre-modern economy is invariably translated into more children, but technology itself depends on the size of population so that at some point the population reaches a "take-off" point where the phase transition can take place. To assure this result, the model postulates that the traditional economy is at "subsistence" at which all increases in income are translated into more children, an extreme version of a Malthusian equilibrium. Once consumption exceeds a threshold, parents start investing in child quality instead of quantity, and the improved education of the next generation then generates more technological change. The idea is conveyed in fig. 3. In panel (a), depicting a "small population" economy, the only equilibrium is at the origin, with no education and no sustained technological progress. As population increases, however, the growth curve g_{t+1} shifts up to create a situation like panel (b) in which there are two stable equilibria (beside the unstable one at $<e_u,g_u>$). They are the low-education ("bad") equilibrium at g_l where education is zero and technological progress is low. It might seem that only a powerful shock can "bump" the economy to the right of the unstable equilibrium to move to the high equilibrium, but keep in mind that if population is growing, the curve g_{t+1} keeps moving up until the economy ends up in a situation like panel (c) where the old "bad equilibrium" is no longer viable and the economy converges to an equilibrium of high education and rapid technological progress.

A third model, in which technology plays a "behind the scenes" role, is the highly original and provocative model by Galor and Moav (2002). In that model, the phase transition is generated by evolutionary forces and natural selection. The idea is that there are two classes of people, those who have many children (r-strategists) and others (K-strategists) who have relatively few but "high-quality" offspring and who invest more in education. Much like Galor and Weil (2000), Lucas (2002), and others, this paper assumes that the world before 1750 was "Malthusian," but it, too, was not entirely static. Because fertility (plus survival) rates were associated with income in the Malthusian economy, the system provided an advantage to K-strategists and their proportion in the population edges up slowly.

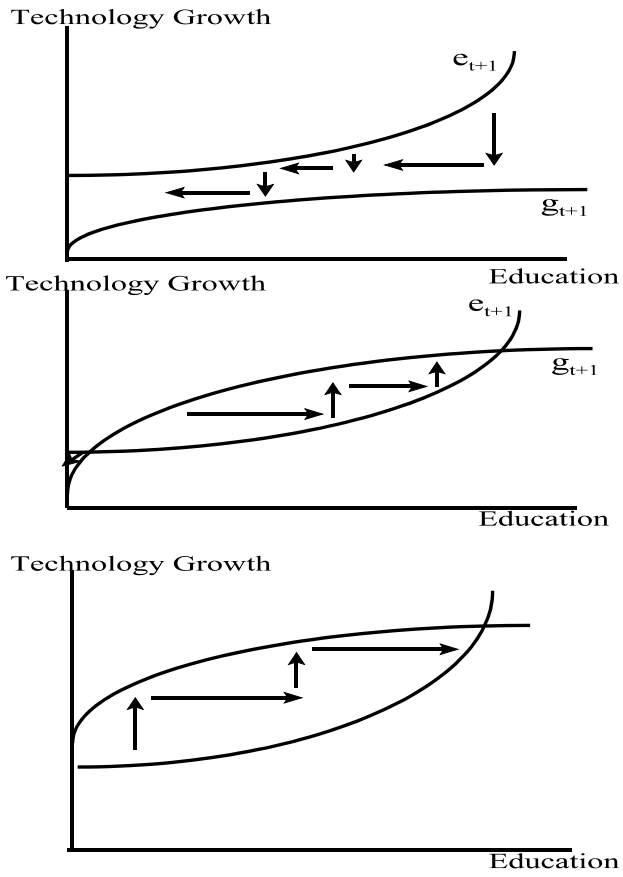
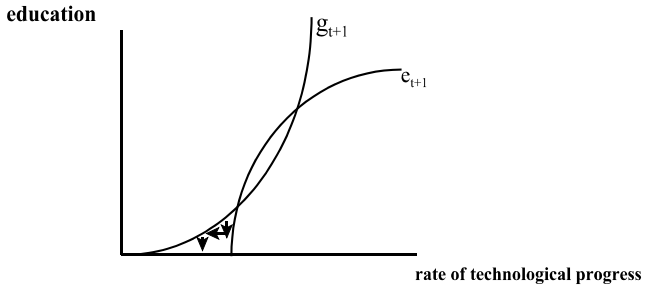


Fig. 3: Transition to a Modern Economy

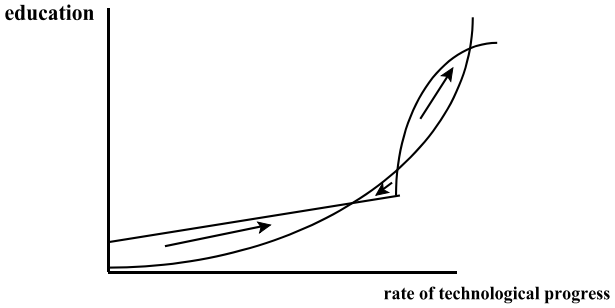
Source: Galor and Weil, 2000

When "quality types" are selected for, more smart and creative people are added and technology advances. Technological progress increases the rate of return to human capital, induces more people to have more "high quality" (educated) children which provides the positive feedback loop. Moreover, as income advances, households have more resources to spend on education, which add to further expansion. Again, technology in this model is wholly endogenous to education and investment in human capital, and an autonomous development in human knowledge and the interplay between propositional and prescriptive knowledge is not really modeled.

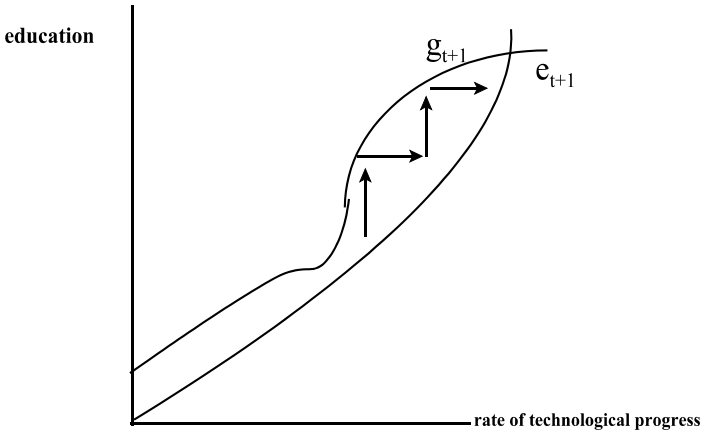
The way the Galor-Moav model reproduces a phase transition is depicted below in fig.4. The diagrams represent a locus of equilibrium points between education at time t (e_t) and growth the next period (g_{t+1}). In panel a, the economy is at an equilibrium at the origin, in the absence of K-strategists who invest in the education of their children. The economy is poor and static, the very picture of a Malthusian state. However, if there are any k-strategists in this economy, their number will steadily grow, causing the e locus to shift to the left, producing an intermediate economy with multiple equilibria. Much like in the Galor-Weil model, this system has a low- and a high-level equilibrium, but because of the changing composition of the economy, the e_{t+1} curve keeps shifting up until the low level steady state $<e_L,g_L>$ is no longer an equilibrium as in panel c. From then on, the economy moves into a steady state with constant growth and a high level of education. An ironic twist of this model that in the end the K-strategists, who are responsible for the transitional dynamics, will be outbred by the r-strategists and vanish. Despite the somewhat limiting assumptions of this model (the "type" is purely inherited and not a choice variable), this paper presents an innovative way of looking at the problem of human capital formation and economic growth in the historical context of the Industrial Revolution.



a: Evolution of Education and Technological progress in the absence of "quality" types



b: Evolution of Education and Technological progress with a small fraction of "quality" types



c. Evolution of Education and Technological progress for a large Fraction of "quality types"

Fig. 4: Education and Technology in an Evolutionary Model.
Source: Galor and Moav (2002), pp. 1163-65

In some sense Galor and Moav's reliance on evolutionary logic to explain technological progress is ironic. In recent years it has been realized increasingly that knowledge *itself* is subject to evolutionary dynamics, in that new ideas and knowledge emerge and are selected for (or not), and that knowledge systems follow a highly path-dependent trajectory governed by Darwinian forces (Ziman, 2000; Mokyr 2003). Yet this insight still awaits to be incorporated in the "take-off" models of growth theorists. Evolutionary models predict that sudden accelerations or "explosions" of evolutionary change (known oddly as "adaptive radiation") occur when conditions are ripe, such as the so-called Cambrian explosion which has been compared to the Industrial Revolution (Kauffman, 1995, p. 205). Another example of rapid evolutionary innovation is the spectacular proliferation of mammals at the beginning of the Cenozoic after the disappearance of the giant reptiles. The idea that evolution proceeds in the highly non-linear rhythm known as "punctuated equilibrium" has been suggested as a possible insight that economic historians can adapt from evolutionary biology (Mokyr, 1990).

Some of these (and other, similar) models may be more realistic than others, and economic historians may have to sort them out. A phase transition model without any reliance on the quality of children and human capital is proposed by Charles Jones (2001) relying on earlier work by Michael Kremer (1993). In Jones's model, what matters is the size rather than the quality of the labor force. In very small populations, the few new ideas that occur lead in straightforward Malthusian fashion to higher populations and not to higher income per capita. As the population gets larger and larger, however, new ideas become more and more frequent, and productivity pulls ahead. The model assumes increasing returns in population and thus generates a classic multiple equilibria kind of story. The positive feedback thus works through fertility behavior responding to higher productivity, and through an increasing returns to population model. As per capita consumption increases, parents substitute away from children to consume other goods, and fertility declines. In this fashion these models succeed in generating both a sudden and discontinuous growth of income per capita or consumption and the fertility transition. Jones shows that for reasonable parameter values he can simulate a world economy that reproduces the broad outlines of modern economic history (including an initial rise in fertility in the early stages of the Industrial Revolution, followed by a decline). Yet the exact connection between the demographic changes and the economic changes in the post 1750 period are far from

understood, and much of the new growth literature pays scant attention to many variables that surely must have affected the demand for children and fertility behavior. These include technological changes in contraceptive technology, a decline in infant- and child mortality, and the changing role of children in the household economy due to the rise in modern technology. It is also open to question whether and to what extent "numbers matter," that is, technology is driven by the flow of "new ideas" and the more people are around, the more likely — all other things equal — new ideas are to emerge.[50] The real question is whether the ideas that count are really a monotonic function of population size (Jones assumes a positive elasticity of .75 to generate his results), or whether they are generated by a minority so small that they are a negligible fraction of the population.[51] The historical record on that is subject to serious debate. It might be added that population growth in Britain was almost nil in the first half of the eighteenth century, and while it took off during the post 1750 era, the same was true for Ireland, where no Industrial Revolution of any kind can be detected.

Most endogenous growth historical models, however, depend on the notion that investment in human capital is critical to the process of "take-off" or phase transition.[52] Historically, however, such a view is not unproblematic either. The idea that the fertility reduction due to changing rates of return on human capital, especially advanced by Lucas (2002), runs into the Fertility Paradox: the first nation to clearly reduce its fertility rate through a decline in marital fertility (that is, intentional and conscious behavior) was *not* the country in which advanced technological techniques were adopted in manufacturing, but France. In Britain fertility rates come down eventually, but the decline does not start until the mid 1870s, a century after the beginning of the Industrial Revolution (e.g., Tranter, 1985, chapter 4).

---

[50] The pedigree of this idea clearly goes back to the work of Julian Simon (1977, 2000).

[51] This sensitivity is reflected in Jones's simulations: the proportion inventors in the population in 1700 in his computations (set to match the demographic data) is 0.875%, but it *declines* in1800 to less than half that number. By constraining the twentieth century data to stay at that level, Jones shows that the Industrial Revolution would be delayed by 300 years.

[52] For a similar view advanced by an economic historian before the new growth economics, see Easterlin (1981).

To argue, therefore, that technological progress was rooted in demographic behavior (through smaller families) seems at variance with the facts. It may well be that in the twentieth century this nexus held, but given the decline in wage premia it is hard to see the rate of return on human capital to be the driving factor.

Moreover, Britain, the most advanced industrial nation in 1850, was far from being the best educated, the most literate, or in some other way the best-endowed in traditional human capital. Increases in male literacy in Britain during the Industrial Revolution were in fact modest and its educational system as a whole lagging behind (Mitch, 1998). The Lutheran nations of the Continent — Germany and the Scandinavian nations — were far more literate and, in one formulation, "impoverished sophisticates."[53] Jewish minorities throughout European history were unusually well-endowed in human capital (Botticini and Eckstein, 2003) yet contributed little or nothing to the Industrial Revolution before 1850. Clearly human capital as a concept is indispensable, but we need to be far more specific as to what *kind* of human capital was produced, for and by whom, what was the source of the demand for it, and how it was distributed over the population. In his recent survey, the social historian Peter Kirby (2003, p. 118) concludes that the idea that nineteenth century education and literacy emerged as a response to a need for a trained labor force is misleading. There was a significant gap between formal 'education' and 'occupational training, the latter remaining embedded in the workplace in the form of apprenticeships and trainee positions. Before 1870, at least, the rate of return on formal education in his view was so low that its benefits did not outweigh the costs. That is not to say that being literate did not convey advantages in terms of social and occupational mobility, but many of the skills that we associate with schooling could be attained informally.

The historical role of human capital in economic growth must be re-examined with some care. In terms of the framework delineated here its importance was first in reducing access costs: literate and educated innovators could and did read articles, books and personal letters from scientists, as well as familiarize themselves with techniques used elsewhere. By knowing more, the cost of verification fell: some obviously bogus and ineffective pieces of propositional knowledge could be rejected offhand.

---

[53] This is a term used by Lars Sandberg in a pathbreaking paper (1979).

Secondly, a more literate and educated labor force could be assumed to be more competent, that is, be able to execute instructions contained in more and more complex techniques. Yet because the process could be divided up more and more, thanks to better access, the actual amount of such knowledge that a *single* worker had to control may not have increased, it may have just changed. Human capital may have been more important in learning new instructions, than in executing more complex and difficult techniques. Above all, human capital made the prepared minds that, as Pasteur famously said, are favored by Fortune. Much technological progress consisted of fumbling and stumbling into some lucky find – but only systematic training allowed inventors to recognize what they found and how to apply it most fruitfully. Yet it is a fair question to ask of all economists who draw links between demographic behavior — whether through the quality-quantity trade-off or otherwise—how many inventors and technically truly able people were needed to generate sustained technological progress.

Eighteenth century Britain did have a cadre of highly skilled technicians and mechanics, most of them trained in the apprenticeship system rather than in formal academies, and these contributed materially to its technological development. But the process of training apprentices did not always correspond to the neoclassical depiction of human capital formation. In addition to imparting skills, it was a selection process in which naturally gifted mechanics would teach themselves from whatever source was available as much as learn from their masters. In the eighteenth century the publishing industry supplied a large flow of popular science books, encyclopedias, technical dictionaries and similar "teach-yourself" kind of books.[54] These mechanics and technicians were the ones that made the Industrial Revolution possible by generating a stream of microinventions that accounted for the actual productivity gains when the great breakthroughs or macroinventions created the opportunities to do so and by providing the competence to carry out the new instructions, that is, to build and construct the new devices according to blueprints and specifications.

Technological change in the era of the Industrial Revolution, based on invention, innovation, and implementation, did not necessarily require

---

[54] Among the many eminent self-educated scientists was Michael Faraday, whose interests in electricity were first stimulated by reading an article in the *Encyclopedia Britannica*.

that the entire labor force or most of it, much less the population at large, be highly educated; that depended entirely on whether the relation between innovation and the growth of competence was strong and positive. An economy that is growing technologically more sophisticated and more productive may end up using techniques that are more difficult to invent and artefacts that are more complex in design and construction, but may be easier to actually use and run on the shop floor. Production techniques became more modular and standardized, meaning that labor might become more specialized and that each worker had to know less rather than more. If much of the new technology introduced after 1825 was like the self-actor—simpler to use if more complex to build— it may well be that the best model to explain technological progress (in the sense of inventing new techniques rather than implementing existing ones) is not the *mean* level of human capital (or, as model-builders have it, the level of human capital of a representative agent), but just the density in the upper tail of the distribution, that is, the level of education and sophistication of a small and pivotal elite of engineers, mechanics, and chemists, dexterous, motivated, imaginative, well-trained technically, with some understanding of some of the science involved. What knowledge the firms could not supply from its own workforce, it purchased from the outside in the form of consulting engineers.[55] Arguably, the system also depended on the increased skills of lower-level technicians, supervisors, foremen, and skilled artisans that the factories needed to introduce new techniques on the shop floor and to make the necessary adjustments to specific tasks and usages. But the bulk of the labor force consisted of rank-and file workers who were in a different category, and thus any model that relates human capital to demographic behavior runs into a serious dilemma. It stands to reason that the ratio of competence to knowledge is higher in agriculture than in manufacturing and in services, since a great deal of competence concerns uncodified knowledge about very local conditions of soil and weather. As the share of agriculture in the labor force and total output declined, this may be one reason why the

[55] Such outside professional consultants included the famous British "coal-viewers" who advised coal mine owners not only on the optimal location and structure of coal mines but also on the use of the Newcomen steam pumps employed in mines in the eighteenth century (Pollard, 1968, pp. 152–53). "Civil engineers" was a term coined by the great engineer John Smeaton, who spent much of his life "consulting" to a large number of customers in need of technical advice.

relative importance of this form of human capital has declined in the twentieth century.

The human capital argument can be tested, at a rudimentary level, by looking at the ratio between skilled and unskilled wages (or wage premium). The problem is of course that without estimating a complete model of the market for skills, the historical course of that ratio cannot be assigned to demand or supply factors. If, however, we assume that technology is the prime mover in this market and we keep in mind that the supply of skills will lag considerably behind a rise in wages (since the acquisition of skills takes time), it would stand to reason that we should observe some increase in the skill premium during the Industrial Revolution. No such change can be observed. Indeed, one of the most surprising facts is that in the twentieth century this skill ratio declined precipitously (Knowles and Robertson, 1951). This could be caused by an (otherwise unexplained) increase in supply, but it is at least consistent with a story that stresses the ability of unskilled labor to operate effectively in a sophisticated technology environment.

The argument I propose, that technological progress is driven by a relatively small number of pivotal people, is not a call for a return to the long-defunct "heroic inventor" interpretation of the Industrial Revolution. The great British inventors stood on the shoulders of those who provided them with the wherewithal of tools and workmanship. John Wilkinson, it is often remarked, was indispensable for the success of James Watt, because his Bradley works had the skilled workers and equipment to bore the cylinders exactly according to specification. Mechanics and instrument makers such as Jesse Ramsden, Edward Nairn, Joseph Bramah, and Henry Maudslay; clock makers such as Henry Hindley, Benjamin Huntsman (the inventor of the crucible technique in making high quality steel), John Whitehurst (a member of the Lunar Society), and John Kay of Warrington (not to be confused with his namesake, the inventor of the flying shuttle, who was trained as a reed and comb maker), engineers such as John Smeaton, Richard Roberts, and Marc I. Brunel; ironmasters such as the Darbys, the Crowleys, and the Crawshays; steam engine specialists such as William Murdoch and Richard Trevithick; chemists such as John Roebuck, Alexander Chisholm, and James Keir were as much part of the story as the "superstars" Arkwright, Cort, Crompton, Hargreaves, Cartwright, Trevithick, and Watt. More often than not, these were men for whom Griffiths's judgment of William Murdoch (the gifted and ingenious Watt and Boulton

employee, credited with the invention of the famous Sun-and-Planets gear) holds: “his inventiveness was instinctive, not analytical. He had an innate sense of mechanical propriety, of the *chose juste*, which led him to simple, robust and highly original solutions” (Griffiths, 1992, p. 209). These were obviously men who could squeeze a great deal out of a narrow epistemic base who could recognize more effective useful knowledge and base better techniques on them. In the end, however, there was no escaping a more formal and analytical approach, in which a widening reliance on physics and mathematics was inevitable. Oddly enough, this approach originated in France more than in Britain.[56]

Below the great engineers came a much larger contingent of skilled artisans and mechanics, upon whose dexterity and adroitness the top inventors and thus Britain's technological success relied. These were the craftsmen who could accurately produce the parts, using the correct dimensions and materials, who could read blueprints and compute velocities, understood tolerance, resistance, and the interdependence of mechanical parts. These were the applied chemists who could manipulate laboratory equipment and acids, the doctors whose advice sometimes saved lives even if nobody quite yet understood why, agricultural specialists who experimented with new breeds of animals, fertilizers, drainage systems, and fodder crops. These anonymous but capable workers produced a cumulative stream of small, incremental, but cumulatively indispensable microinventions without which Britain would not have become the "workshop of the world." It is perhaps premature to speak of an "invention industry" by this period, but technical knowledge at a level beyond the reach of the run-of-the-mill artisan became increasingly essential to create the inventions associated with the Industrial Revolution.[57]

---

[56] The “Big Three polytechnicien” engineers of the early nineteenth century, Gustave-Gaspard Coriolis, Jean-Victor Poncelet, and Louis Navier, placed mechanical and civil engineering on a formal base, and supported practical ideas with more formal theory than their more pragmatic British colleagues

[57] A number of high-skill sectors that had developed in Britain since 1650 played important roles in them. Among those instrument- and clock making, mining, and ship yards are of central importance. Cardwell (1972, p. 74) points out that a number of basic technologies converge on mining (chemistry, civil engineering, metallurgy) and that mining sets the hard, "man-sized" problems, controlling powerful forces of nature and transforming materials on a large scale. In addition, however,

The average "quality" of the rest of the labor force – in terms of their technical training – may have mattered to the development and adoption of the new techniques less than is commonly believed. A venerable tradition in economic history, in fact, has argued that technological progress in the first stages of the Industrial Revolution was "deskilling," requiring workers who were able to carry out repetitive routine actions instead of the skilled labor of skilled craftsmen.[58] The "factory system" required workers to be supervised and assisted by skilled mechanics, and hence the variance of the skill level may have increased even if we cannot be sure whether *average* skills had to go up or down. Moreover, the term "skill" may be too confining. Human capital was in part produced in schools, but what future workers were taught in schools may have had as much to do with behavior as with competence. Docility and punctuality were important characteristics that factory owners expected from their workers. "The concept of industrial discipline was new, and called for as much innovation as the technical inventions of the age," writes Pollard ([1965] 1968, p. 217). Early factories designed incentives to bring about the discipline, but they also preferred to hire women and children, who were believed to be more docile. Some of the literature by economists on human capital acquisition may have to be reinterpreted in this fashion. Beyond that, however, human capital was instrumental in creating competence rather than knowledge itself. Yet given that much of competence consisted of tacit knowledge and experience, and given that much of the competence could be front-loaded into the equipment by a small number of brilliant designers, the role of either the size of the population or their "mean" level of human capital should be questioned. It seems plausible that the degree of networking and the level of access costs *within* the relatively small community of highly trained engineers and scientists may have been of greater importance.

British millwrights were technologically sophisticated: the engineer John Fairbairn, a millwright himself, noted that eighteenth century British millwrights were "men of superior attainments and intellectual power," and that the typical millwright would have been "a fair arithmetician, knew something of geometry, levelling and mensuration and possessed a very competent knowledge of practical mechanics" (cited in Musson and Robinson, 1969, p. 73).

[58] On the eve of the Industrial Revolution, much of the manufacturing sector in Europe had already been de-skilled to a considerable extent, with much of the production carried out in the homes of unskilled rural workers.

To understand the "phase transition" in the dynamic of useful knowledge, we need to look again at the relationship between propositional and prescriptive knowledge. As the two forms of knowledge co-evolved, they increasingly enriched one another, eventually tipping the balance of the feedback mechanism from negative to positive and creating the phase transition. During the early stages of the Industrial Revolution propositional knowledge mapped into new techniques creating what we call "inventions." This mapping should not be confused with the linear models of science and technology, popular in the mid-twentieth century, which depicted a neat flow from theory to applied science to engineering and from there to technology. Much of the propositional knowledge that led to invention in eighteenth century was artisanal: pragmatic, informal, intuitive, and empirical. Only very gradually did it the kind of formal and consensual knowledge we think of today as "science"become a large component of it. It was, in all cases, a small fraction of what is known today. What matters is that it was subject to endogenous expansion: prescriptive knowledge in its turn enhanced propositional knowledge, and thus provided positive feedback between the two types of knowledge, leading to continuous mutual reinforcement. When powerful enough, this mechanism can account for the change in stability of the entire system. The positive feedback from prescriptive to propositional knowledge took a variety of forms.

One of those forms is what Rosenberg has called "focusing devices:" technology posed certain riddles that science was unable to solve, such as "why (and how) does this technique work." The most celebrated example of such a loop is the connection between steam power and thermodynamics, exemplified in the well–known tale of Sadi Carnot's early formulation, in 1824, of the Second Law of Thermodynamics by watching the difference in fuel economy between a high pressure (Woolf) steam engine and a low pressure one of the Watt type.[59] The next big step was made by an Englishman, James P. Joule, who showed the conversion rates from work to

[59] It is interesting to note that Carnot's now famous *Reflexions sur la puissance motrice du feu* (1824) was initially ignored in France and eventually found its way second hand and through translation into Britain, where there was considerably more interest in his work because of the growing demand by builders of gigantic steam engines such as William Fairbairn in Manchester and Robert Napier in Glasgow for theoretical insights that would help in making better engines.

heat and back.[60] Joule's work and that of Carnot were then reconciled by a German, R. J. E. Clausius (the discoverer of entropy), and by 1850 a new branch of science dubbed "thermodynamics" by William Thomson (later Lord Kelvin) had emerged (Cardwell, 1971, 1994).[61] Power technology and classical energy physics subsequently developed hand-in-hand, culminating in the career of the Scottish physicist and engineer William Rankine whose *Manual of the Steam Engine* (1859) made thermodynamics accessible to engineers and led to a host of improvements in actual engines. In steam power, then, the positive feedback can be clearly traced: the first engines had emerged in the practical world of skilled blacksmiths, millwrights, and instrument makers with only a minimum of theoretical understanding. These machines then inspired theorists to come to grips with the natural regularities at work and to widen the epistemic base. The insights generated were in turn fed back to engineers to construct more efficient engines. This kind of mutually reinforcing process can be identified, in a growing number of activities, throughout the nineteenth century. They required the kind of intellectual environment that the Industrial Enlightenment had created: a world in which technical knowledge was accessible and communicable in an international elite community, a technological invisible college that encompassed much of the Western world.

A less well known example of this feedback mechanism, but equally important to economic welfare, is the interaction between the techniques of food-canning and the evolution of bacteriology. As noted earlier, the canning

---

[60] The ways in which the growth of practical knowledge can influence the emergence of propositional knowledge are well illustrated by Joule's career: he was a child of industrial Lancashire (his father owned a brewery) and in the words of one historian, "with his hard-headed upbringing in industrial Manchester, was unambiguously concerned with the *economic* efficiency of electromagnetic engines...he quite explicitly adopted the language and concerns of the economist and the engineer" (Morus, 1998, p. 187, emphasis in original). As Ziman remarks (1976, p. 26), the first law of thermodynamics could easily have been derived from Newton's dynamics by mathematicians such as Laplace or Lagrange, but it took the cost accountancy of engineers to bring it to light.

[61] Research combining experiment and theory in thermodynamics continued for many decades after that, especially in Scotland and in Mulhouse, France, where Gustave Adolphe Hirn, a textile manufacturer, led a group of scientists in tests on the steam engines in his factory and was able to demonstrate the law of conservation of energy.

of food was invented in 1795, by Nicolas Appert. He discovered that when he placed food in champagne bottles, corked them loosely, immersed them in boiling water, and then hammered the corks tight, the food was preserved for extended periods. Neither Appert nor his English emulators who perfected the preservation of food in tin-plated canisters in 1810 really understood why and how this technique worked, because the definitive demonstration of the notion that microorganisms were responsible for putrefaction of food was still in the future. It is therefore a typical example of a technique with a narrow epistemic base. The canning of food led to a prolonged scientific debate about what caused food to spoil. The debate was not put to rest until Pasteur's work in the early 1860s. Pasteur knew of Appert's work, and eventually admitted that his own work on the preservation of wine was only a new application of Appert's method. Be that as it may, his work on the impossibility of spontaneous generation clearly settled the question of why the technique worked and provided the epistemic base for the technique in use. When the epistemic base of food-canning became wider, techniques improved: the optimal temperatures for the preservation of various foods with minimal damage to flavor and texture were worked out by two MIT scientists, Samuel Prescott and William Underwood.[62]

A different feedback mechanism from prescriptive to propositional knowledge was described by Derek Price as "Artificial Revelation." The idea is fairly simple: our senses limit us to a fairly narrow slice of the universe that has been called a "mesocosm": we cannot see things that are too far away, too small, or not in the visible light spectrum (Wuketits, 1990, pp. 92, 105). The same is true for our other senses, for the ability to make very accurate measurements, for overcoming optical and other sensory illusions, and — perhaps most important in our own time — the computational ability of our brains. Technology consists in part in helping us overcome these limitations that evolution has placed on us and learn of natural phenomena we were not meant to see or hear.[63] The period of the Industrial

---

[62] A University of Wisconsin scientist, H. L. Russell, proposed to increase the temperature of processing peas from 232° to 242°, thus reducing the percentage spoiled can from 5 percent to 0.07 percent (Thorne, 1986, p. 145).

[63] Derek Price notes that Galileo's discovery of the moons of Jupiter was the first time in history that somebody made a discovery that had been totally unavailable to others by a process that did not involve a deep and clever thought (1984b, p. 54).

Revolution witnessed a great deal of improvement in techniques whose purpose it was to enhance propositional knowledge. Lavoisier and his circle designed and used better laboratory equipment that allowed them to carry out more sophisticated experiments.[64] Alessandro Volta invented a pile of alternating silver and zinc disks that could generate an electric current in 1800. Volta's battery was soon produced in industrial quantities by William Cruickshank. Through the new tool of electrolysis, pioneered by Humphry Davy, chemists were able to isolate element after element and fill in much of the detail in the maps whose rough contours had been sketched by Lavoisier and Dalton. Volta's pile, as Davy put it, acted as an "alarm bell to experimenters in every part of Europe" (cited by Brock, 1992, p. 147).Or consider the development of the technique of in vitro culture of micro-organisms (the Petri dish was invented in 1887 by R. J. Petri, an assistant of Koch's). Price feels that such advances in knowledge are "adventitious"(1984a, p. 112). Travis (1989) has documented in detail the connection between the tools developed in the organic chemical industry and advances in cell biology. These connections between prescriptive and propositional knowledge are just a few examples of advances in scientific techniques that can be seen as adaptations of ideas originally meant to serve an entirely different purpose, and they reinforce the contingent and accidental nature of much technological progress (Rosenberg, 1994, pp. 251–52).

The invention of the modern compound microscope attributed to Joseph J. Lister (father of the famous surgeon) in 1830 serves as another good example. Lister was an amateur optician, whose revolutionary method of grinding lenses greatly improved image resolution by eliminating spherical aberrations.[65] His invention and the work of, among others, Pierre Guinand

---

[64] The famous mathematician Pierre-Simon de Laplace was also a skilled designer of equipment and helped to build the calorimeter that resulted in the celebrated "Memoir on Heat" jointly written by Laplace and Lavoisier (in 1783), in which respiration was identified as analogous to burning. Much of the late eighteenth-century chemical revolution was made possible by new instruments such as Volta's eudiometer, a glass container with two electrodes intended to measure the content of air, used by Cavendish to show the nature of water as a compound.

[65] The invention was based on a mathematical optimization for combining lenses to minimize spherical aberration and reduced average image distortion by a huge proportion, from 19 to 3 percent. Lister is reputed to have been the first human being ever to see a red blood cell.

and Samuel Klingenstierna changed microscopy from an amusing diversion to a serious scientific endeavor and eventually allowed Pasteur, Koch, and their disciples to refute spontaneous generation and to establish the germ theory, a topic I return to below. The germ theory was one of the most revolutionary changes in useful knowledge in human history and mapped into a large number of new techniques in medicine, both preventive and clinical. Indeed, the widespread use of glass in lenses and instruments in the West was itself something coincidental, a "giant accident," possibly a by-product of demand for wine and different construction technology (Macfarlane and Martin, 2002). It seems plausible that without access to this rather unique material, the development of propositional knowledge in the West would have taken a different course.[66]

A third mechanism of technology feeding back into prescriptive knowledge is through what might be called the "rhetoric of knowledge." This harks back to the idea of "tightness" introduced earlier. Techniques are not "true" or "false." Either they work as expected or they do not, and thus they can be interpreted to confirm or refute the propositional knowledge that serves as their epistemic base. Prescriptive knowledge has varying degrees of tightness, depending on the degree to which the available evidence squares with the rhetorical conventions for acceptance. Laboratory technology transforms conjecture and hypothesis into an accepted fact, ready to go into textbooks and to be utilized by engineers, physicians, or farmers. But a piece of propositional knowledge in the past was often tested simply by verifying that the techniques based on it actually worked. The earthenware manufacturer Josiah Wedgwood felt that his experiments in pottery actually tested the theories of his friend Joseph Priestley, and professional chemists, including Lavoisier, asked him for advice. Similarly, once biologists discovered that insects could be the vectors of pathogenic microparasites, insect-fighting techniques gained wide acceptance. The success of these techniques in eradicating yellow fever and malaria was the best confirmation of the hypotheses about the transmission mechanisms of the disease and helped earn them wide support.

---

[66] MacFarlane and Martin (2002, pp. 81-82) note that glass lenses not only made possible specific discoveries but led to a growing confidence in a world of deeper truths to be discovered, destabilizing conventional views. "The obvious was no longer true. Hidden connections and buried forces could be analyzed."

Or consider the question of heavier-than-air flight. Much of the knowledge in aeronautics in the early days was experimental rather than theoretical, such as attempts to tabulate coefficients of lift and drag for each wing shape at each angle. It might be added that the epistemic base supporting the first experiments of the Wright brothers was quite untight: in 1901 the astronomer and mathematician Simon Newcomb (the first American since Benjamin Franklin to be elected to the Institute of France) opined that flight carrying anything more than "an insect" would be impossible.[67] The success at Kitty Hawk persuaded all but the most stubborn doubting Thomases that human flight in heavier-than-air fixed wing machines was possible. Clearly their success subsequently inspired a great deal of research on aerodynamics. In 1918 Ludwig Prandtl published his magisterial work on how wings could be scientifically rather than empirically designed and the lift and drag precisely calculated (Constant, 1980, p. 105; Vincenti, 1990, pp. 120–25). Even after Prandtl, not all advances in airplane design were neatly derived from first principles in an epistemic base in aerodynamic theory, and the ancient method of trial and error was still widely used in the search for the best use of flush riveting in holding together the body of the plane or the best way to design landing gear (Vincenti, 1990, pp. 170–99; Vincenti, 2000).

It is important not to exaggerate the speed and abruptness of the transition. Thomas Edison, a paradigmatic inventor of the 2nd Industrial Revolution barely knew any science, and in many ways should be regarded an old-fashioned inventor who invented by trial-and-error through intuition, dexterity and luck. Yet he knew enough to know what he did not know, and that there were others who knew what he needed. Among those who supplied him with the propositonal knowledge necessary for his research were the mathematical physicist Francis Upton, the trained electrical engineer Hermann Claudius, the inventor and engineer Nikola Tesla, the physicist Arthur E. Kennelly (later professor of electrical engineering at Harvard), and the chemist Jonas W. Aylsworth. Yet by that time access costs had declined enough so that he could learn for instance of

---

[67] He was joined in that verdict by the Navy's chief engineer, Admiral George Melville (Kelly, 1943, pp. 116–17; Crouch, 1989, p. 137). Nor were the inventors themselves all that certain: in a widely quoted remark, Wilbur Wright in a despondent mood remarked to his brother that "not within a thousand years would men ever fly" (Kelly, 1943, p. 72).

the work of the great German physicist Hermann von Helmholtz through a translated copy of the latter's work on acoustics.

The positive feedback from technology to prescriptive knowledge entered a new era with development of the computer. In the past, the practical difficulty of solving differential equations limited the application of theoretical models to engineering. A clever physicist, it has been said, is somebody who can rearrange the parameters of an insoluble equation so that it does not have to be solved. Computer simulation can evade that difficulty and help us see relations in the absence of exact closed-form solutions and may represent the ultimate example of Bacon's "vexing" of nature. In recent years simulation models have been extended to include the effects of chemical compounds on human bodies. Combinatorial chemistry and molecular biology are both equally unimaginable without fast computers. It is easy to see how the mutual reinforcement of computers and their epistemic base can produce a virtuous circle that spirals uncontrollably away from its basin of attraction. Such instability is the hallmark of Kuznets's vision of the role of "useful knowledge" in economic growth.

In addition to the positive feedback within the two types of knowledge, one might add the obvious observation that *access costs* were themselves a function of improving techniques, through better communications, storage, and travel techniques. In this fashion, expansions in prescriptive knowledge not only expanded the underlying supporting knowledge but made it more accessible and thus more likely to be used. As already noted, this is particularly important because so much technological progress consists of combinations and applications of existing techniques in novel ways, or parallels from other techniques in use. Precisely for this reason, cheap and reliable access to the monster catalog of all feasible techniques is an important element in technological progress. As the total body of useful knowledge is expanding dramatically in our own time, it is only with the help of increasingly sophisticated search engines that needles of useful knowledge can be retrieved from a haystack of cosmic magnitude.

Technological modernity is created when the positive feedback from the two types of knowledge becomes self-reinforcing and autocatalytic. We could think of this as a phase transition in economic history, in which the old parameters no longer hold, and in which the system's dynamics have been unalterably changed. There is no necessity for this to be true even in the presence of positive feedback; but for certain levels of the parameters, the system as a whole becomes unstable. It may well that this instability in the

knowledge-producing system are what is behind what we think of as "technological modernity." Kuznets, of course, felt that the essence of modern growth was the increasing reliance of technology on modern science. This view, as I have argued above needs clarification and amplification. Inside the black box of technology is a smaller black box called "research and development" which translates inputs into the output of knowledge. This black box itself contains an even smaller black box which models the available knowledge in society, and it is this last box I have tried to pry open. Yet all this is only part of the story: knowledge creates opportunities, but it does not guarantee action. Knowledge is an abstract concept, it glosses over the human agents who possess it and decide to act upon it. What motivates them, and why did some societies seem to be so much more inclined to generate new knowledge and to exploit the knowledge it had? To understand why during the past two centuries the "West" has been able to take advantage of these opportunities we need to examine the institutional context of innovation.

Institutions and Technological progress

Beyond the interaction of different kinds of knowledge was the further level of interaction and feedback between human knowledge and the institutional environment in which it operates. Before 1750, economic progress of any kind had tended to run into what could best be called negative institutional feedback. One of the few reliable regularities of the premodern world was that whenever a society managed, through thrift, enterprise, or ingenuity to raise its standard of living, a variety of opportunistic parasites and predators were always ready to use power, influence, and violence to appropriate this wealth. Such rent-seekers, who redistributed wealth rather than created it, came either from within the economy in the form of tax-collectors, exclusive coalitions, and thugs, or they came from outside as alien pillagers, mercenaries, and plunderers. The most obvious and costly form of negative institutional feedback before 1815 was, of course, war. Rent-seeking and war often went in hand in hand. Britain, France, the United Provinces and most other Continental powers fought one another constantly in hugely costly attempts to redistribute taxable real estate, citizens and activities from one to the other, a typical "mercantilist" kind of

policy.[68] Economic growth indirectly helped instigate these conflicts. Wealth accumulation, precisely because it was mostly the result of "Smithian Growth," was usually confined to a region or city and thus created an incentive to greedy and well-armed neighbors to engage in armed rent-seeking. It surely is no accident that the only areas that had been able to thwart off such marauders with some success were those with natural defenses such as Britain and the Netherlands. Yet the Dutch United Provinces were weakened by the relentless aggressive mercantilist policies of powerful neighbors.[69] The riches of the Southern Netherlands – unfortunately easier to invade – were repeatedly laid to waste by invading mercenary soldiers after 1570. More subtle forms of rent-seeking came from local monopolists (whose claims to a right to exclude others were often purchased from strongmen), guilds with exclusionary rights, or nobles with traditional rights such as *banalités*. A particularly harmful form of rent-seeking were price controls on grain that redistributed resources from the countryside to the city by keeping grain prices at below equilibrium levels (Root, 1994).

Had institutional feedback remained negative, as it had been before 1750, the economic benefits of technological progress would have remained limited. Mercantilism, as Ekelund and Tollison (1981, 1997) have emphasized, was largely a system of rent seeking, in which powerful political institutions redistributed wealth from foreigners to themselves as well between different groups and individuals within the society. The Industrial Enlightenment meant that the old rent-seeking traditions of exclusionary privileges were increasingly viewed as both unfair and inefficient.

---

[68] O'Brien (2003, p. 5) notes that between the nine-years war (starting in 1688) and the Congress of Vienna in 1815, Britain and France were at or on the brink of war for more than half the period, justifying the term "Second Hundred Years War."

[69] The standard argument is that national defense was so costly that high indirect taxes led to high nominal wages, which rendered much of Dutch manufacturing uncompetitive. See for example Charles Wilson (1969). De Vries and Van Der Woude (1997, p. 680) point out that in 1688 the Dutch committed huge resources to an invasion of England because the future economic well-being on the Republic depended on the destruction of French mercantilism and the establishment of an international order in which the Dutch economy could prosper, yet it "proved to be a profitless investment." More recently, Ormrod (2003) has confirmed the view that the decline of the Dutch Republic was a direct consequence of the mercantilist policies of its neighbors, especially Britain.

Mercantilism had been a game of international competition between rival political entities. To defeat an opponent, a nation had to outcompete it, which it often did by subsidizing exports and raw materials imports, and imposing a tariff on finished goods. As it dawned upon people that higher productivity could equally outcompete other producers, they switched to a different policy regime, one that economists certainly would recognize as more enlightened.[70] In the century before 1750, mercantilism had begun to decline in certain key regions in Western Europe, above all in Britain, where many redistributive arrangements such as guilds, monopolies, and grain price regulations were gradually weakening, though their formal disappearance was still largely in the future. The Age of Enlightenment led to some pre-1789 reforms on the Continent thanks to a few enlightened despots, but it seems beyond doubt that the French Revolution and the ensuing political turmoil did more than anything else to transform Enlightenment ideas into institutional changes that paved the road for economic growth (Mokyr, 2003b). The Enlightenment also advocated more harmonious and cosmopolitan altitudes in international relations and its influence may have contributed to the relative calm that settled upon Europe after the Congress of Vienna. Political reforms that weakened privileges and permitted the emergence of freer and more competitive markets had an important effect on economic performance. The institutional changes in the years between 1770 and 1815 saw to it that the Industrial Revolution was not followed by a surge in rent-seeking and violence that eventually could have reversed the process.

The feedback between technological and institutional change is central to the process of historical change. The co-evolution of technological knowledge and institutions during the second Industrial Revolution has been noticed before.[71] Above all, three kind of institutions were important in

---

[70] In 1773, the steam engine manufacturer Matthew Boulton told Lord Harwich that mechanization and specialization made it possible for Birmingham manufacturers to defeat their Continental competitors (cited by Uglow, 2002, p. 212).

[71] Nelson (1994) has pointed to a classic example, namely the growth of the large American business corporation in the closing decades of the nineteenth century, which evolved jointly with the high-throughput technology of mass production and continuous flow. In their pathbreaking book, Fox and Guagnini (1999) have pointed to the growth of practically-minded research laboratories in academic communities, which increasingly cooperated and interacted successfully with industrial establishments to

facilitating the sustained technological progress central to economic growth: those that provided for connections between the people concerned mostly with propositional knowledge and those on the production side; those that set the agenda of research to generate new propositional knowledge that could be mapped into new techniques; and those institutions that created *incentives* for innovative people to actually spend resources in order to map this knowledge into actual techniques and specifically that weakened the effective social and political resistance against new techniques. As noted above, even some of the formal endogenous growth models require a growing proportion of labor in the "invention sector," a condition that clearly demands that their profits not be expropriated altogether.

The institutions that created the bridges between prescriptive and propositional knowledge in late eighteenth and nineteenth century Europe are well understood: universities, polytechnic schools, publicly funded research institutes, museums, agricultural research stations, research departments in large financial institutions. Improved access to useful knowledge took many forms: cheap and widely diffused publications disseminated it. Technical subjects penetrated school curricula in every country in the West (although Britain, the leader in the first Industrial Revolution, lost its momentum in the last decades of the Victorian era). All over the Western world, textbooks of applied science (or "experimental philosophy" in the odd terminology of the time), professional journals, technical encyclopedias, and engineering manuals appeared in every field and made it easier to "look things up." The professionalization of expertise meant increasingly that anyone who needed some piece of useful knowledge could find someone who knew, or who knew who knew. The learned journal first appeared in the 1660s and by the late eighteenth century had become one of the main vehicles by which prescriptive knowledge was diffused. In the eighteenth century, most scientific journals were in fact deliberately made accessible, because they more often than not catered to a lay audience and thus were media of education and dissemination rather than repositories of original contributions (Kronick, 1962, p. 104). Review articles and book

---

create an ever-growing stream of technological adaptations and microinventions. Many other examples can be cited, such as the miraculous expansion of the British capital market which emerged jointly with the capital-hungry early railroads and the changes in municipal management resulting from the growing realization of the impact of sanitation on public health (Cain and Rotella, 2001).

reviews that summarized and abstracted books and learned papers (especially those published overseas and were less accessible), another obvious example of an access-cost reduction, were popular.[72] In the nineteenth century, specialized scientific journals became increasingly common and further reduced access costs, if perhaps more and more requiring the intermediation of experts who could decode the jargon.

To be sure, co-evolution did not always quickly produce the desired results. British engineering found it difficult to train engineers using best-practice knowledge, and the connections between science and engineering remained looser and weaker than elsewhere. In 1870 a panel appointed by the Institute of Civil Engineers concluded that "the education of an Engineer is effected by...a simple course of apprenticeship to a practicing engineer...it is not the custom in England to consider *theoretical* knowledge as absolutely essential" (cited by Buchanan, 1985, p. 225). A few individuals, above all William Rankine at Glasgow, argued forcefully for more bridges between theory and practice, but significantly he dropped his membership in the Institute of Civil Engineers. Only in the late nineteenth century did engineering become a respected discipline in British universities.

Elsewhere in Europe, the emergence of universities and technical colleges that combined research and teaching, thus simultaneously expanding propositional knowledge and reducing access costs, advanced rapidly. An especially good and persuasive example is provided by Murmann (1998), who describes the co-evolution of technology and institutions in the chemical industry in imperial Germany, where the new technology of dyes, explosives, and fertilizers emerged in constant interaction with the growth of research and development facilities, institutes of higher education, and large industrial corporations with a knack for industrial research.[73] Institutions remained a major determinant of access costs. To understand the

---

[72] This aspect of the Industrial Enlightenment was personified by the Scottish writer and mathematician John Playfair (1748-1819) whose textbooks and review essays in the *Edinburgh Review* made a special effort to incorporate the work of Continental mathematicians, as witnessed by the essays in 1807 on the work of Mechain and Delambre on the earth's meridian, and his 1808 review of LaPlace's *Traité de Mécanique Celeste* (Chitnis, 1976, pp. 176-77, 222).

[73] Most famous, perhaps, was the invention of alizarin in 1869, a result of the collaboration between the research director at BASF, Caro, with the two academics Graebe and Liebermann.

evolution of knowledge, we need to ask who talked to whom and who read what. Yet the German example illustrates that progress in this area was halting and complex; it needs to be treated with caution as a causal factor in explaining systematic differences between nations. The famed *technische Hochschulen*, the German equivalent of the French *polytechniques,* had lower social prestige than the universities and were not allowed to award engineering diplomas and doctorates till 1899. The same is true for the practical, technically oriented *Realschulen* which had lower standing than the more classically inclined *Gymnasien*. Universities conducted a great deal of research, but it goes too far to state that what they did was a *deliberate* application of science to business problems.[74] Universities and businesses co-evolved, collaborating through intense communications, overlapping personnel, and revolving doors. The second Industrial Revolution rested as much on industry-based science as on the more common concept of science-based industry (König, 1996).

Designing institutions that create the correct ex ante motivations to encourage invention is not an easy task. Economists typically believe that agents respond to economic incentives. A system of relatively secure property rights, such as emerged in Britain in the seventeenth century, clearly was prerequisite. Without it, even if useful knowledge would expand, the investment and entrepreneurship required for a large scale implementation of the new knowledge would not have been forthcoming. On a more specific level, the question of intellectual property rights and rewards for those who add to the stock of useful knowledge is paramount. Some of the best recent work in the economic history of technological change focuses on the working of the patent system as a way of preserving property rights for inventors. In a series of ingenious papers, Kenneth Sokoloff and Zorina Khan have shown how the American patent system exhibited many of the characteristics of a market system: inventors responded to demand conditions, did all they could to secure the gains from

---

[74] James (1990, p. 111) argues that Germany's "staggering supremacy" was not due to scientists looking for applicable results but came about "because her scientists experimented widely without any end in mind and then discovered that they could apply their new information." This seems a little overstated, but all the same we should be cautious in attributing too much intent and directionality in the growth of knowledge. Much of it was in part random or the unintended consequence of a different activity, and it was the selection process that gave it its technological significance. In that respect, the evolutionary nature of the growth in useful knowledge is reaffirmed.

their invention and bought and sold licenses in what appears to be a rational fashion. It was far more accessible, open, and cheaper to use than the British system, and attracted ordinary artisans and farmer as much as it did professional inventors and eccentrics (Khan and Sokoloff, 1993, 1998, 2001; Khan, 2002).

Whether this difference demonstrates that a well-functioning system of intellectual property rights is truly essential to the growth of useful knowledge remains an open question. For one thing, the American system was far more user-friendly than the British patent system prior to its reform in 1852. Yet despite the obvious superiority of the U.S. system and the consequent higher propensity of Americans to patent, there can be little doubt that the period between 1791 and 1850 coincides roughly with the apex of British superiority in invention. The period of growing American technological leadership, after 1900, witnessed a stagnation and then a decline in the American per capita patenting rate. Other means of appropriating the returns on R&D became relatively more attractive. In Britain, MacLeod (1988) has shown that the patent system provided only weak and erratic protection to inventors and that large areas of innovation were not patentable. Patenting was associated with commercialization and the rise of a profit-oriented spirit, but its exact relation to technological progress is still obscure.[75] What is sometimes overlooked is that patents placed technical information in the public realm and thus reduced access costs. Inventors, by observing what had been done, saw what was possible and were inspired to apply the knowledge thus acquired to other areas not covered by the patent. In the United States, *Scientific American* published lists of new patents from 1845, and these lists were widely consulted. Despite the limitations that patents imposed on applications, they reduced access costs to the knowledge

[75] In fact, economists have argued that for countries that are technologically relatively backward, strict patent systems may be on balance detrimental to economic welfare (for a summary, see Lerner, 2000). In a different context, Hilaire-Pérez (2000) has shown how different systems of invention encouragement in eighteenth-century Europe were consistent with inventive activity: whereas in France the state played an active role of awarding "privileges" and pensions to inventors deemed worthy by the French Academy, in Britain the state was more passive and allowed the market to determine the rewards of a successful inventor. These systems were not consistently enforced (some British inventors whose patents for one reason or another failed to pay off were compensated by special dispensation) and, as Hilaire-Pérez shows, influenced one another.

embodied in them. This function of the patent system apparently was fully realized in the 1770s. The full specification of patents was meant to inform the public. In Britain this was laid out in a decision by chief justice Lord Mansfield, who decreed in 1778 that the specifications should be sufficiently precise and detailed so as to fully explain it to a technically educated person. In the Netherlands, where patenting had existed from the 1580s, the practice of specification was abandoned in the mid-1630s but revived in the 1770s (Davids, 2000, p. 267).

In at least two countries, the Netherlands and Switzerland, the complete absence of a patent system in the second half of the nineteenth century does not seem to have affected the rate of technological advance (Schiff, 1971). Of course, being small, such countries could and did free- ride on technological advances made elsewhere, and it would be a fallacy to infer from the Dutch and Swiss experience that patents did not matter. It also seems plausible that reverse causation explains part of what association there was between the propensity to patent and the generation of new techniques: countries in which there were strong and accessible bridges between the *savants* and the *fabricants* would feel relatively more need to protect the offspring of these contacts. Lerner (2000) has shown that rich and democratic economies, on the whole, provided more extensive patent protection. The causal chain could thus run from technological success to income and from there to institutional change rather than from the institutions to technological success, as Khan and Sokoloff believe. It may well be true, as Abraham Lincoln said, that what the patent system did was "to add the fuel of interest to the fire of genius" (cited by Khan and Sokoloff, 2001, p. 12), but that reinforces the idea that we need to be able to say something about how the fire got started in the first place.

Other institutions have been widely recognized as aiding in the generation of new techniques. Among those are relatively easy entry and exit from industries, the availability of venture capital in some form, the reduction of uncertainty by a large source of assured demand for a new product or technique (such as military procurement), the existence of agencies that coordinate and standardize the networked components of new techniques, and revolving doors between industry and organizations that specialize in the generation of propositional knowledge such as universities and research institutes.

Before these institutions and the inventions they stimulated, however, had to be the propositional knowledge on which the inventions

rested. Augmenting this knowledge opened the door that economic incentives and markets pushed societies through. Had the doors remained closed, however, improved incentives for innovation wold have been useless. Commercial, entrepreneurial, and even sophisticated capitalist societies have existed that made few important technical advances, simply because the techniques they employed rested on narrow epistemic base and the propositional knowledge from which these bases were drawn was not expanding. The reasons for this could be many: the agendas of intellectual activity may not have placed a high priority on useful knowledge, or a dominant conservative religious philosophy might have stifled a rebellious attitude toward existing propositional knowledge. Above all, there has to be a belief that such knowledge will eventually be socially useful even if the gains are likely to be reaped mostly by persons others than the ones who generate the novel propositional knowledge. Given that increasing this knowledge was costly and often regarded as socially disruptive, the political will by agents who controlled resources to support this endeavor,whether they were rich aristocratic patrons or middle-class taxpayers, was not invariably there. The amounts of resources expended on R&D, however, are not more important than questions about how they are spent, on what, and what kind of access potential users have to this knowledge.

One specific example of an area in which technological innovation and institutional change interacted in this fashion was in the resistance of vested interests to new technology (Mokyr, 1994, 2002). Here institutions are particularly important, because by definitions such resistance has to operate outside of the market mechanism. If left to markets to decide, it seems likely that superior techniques and products will inexorably drive out existing ones. For the technological status quo to fight back thus meant to use non-market mechanisms. These could be legal, through the manipulation of the existing power structure, or extralegal, through machine-breaking, riots, and the use of personal violence against inventors and the entrepreneurs who tried to adopt their inventions.

At one level, eighteenth-century Enlightenment thinking viewed technological change as "progress" and implicitly felt that social resistance to it was socially undesirable. Yet there was a strand of thought, associated with Rousseau and with later elements of romanticism such as Cobbett and Carlyle continuing with the Frankfurt school in the twentieth century, that viewed industrialization and modern technology sincerely as evil and destructive. Such ideological qualm often found themselves allied with those

whose human and physical capital was jeopardized by new techniques. The ensuing battle came to a crashing crescendo during the Industrial Revolution. The Luddite rebellion — a complex set of events that involved a variety of grievances, not all of which were related to rent-seeking – was mercilessly suppressed. It would be a stretch to associate the harsh actions of the British army in the midlands with anything like the Enlightenment. All the same, it appears that rent-seeking inspired resistance against new technology had been driven into a corner by that time by people who believed that "freedom" included the freedom to innovate.

The British example is quite telling.[76] In the textile industries, by far the most resistance occurred in the woolen industries. Cotton was a relatively small industry on the eve of the Industrial Revolution and had weakly entrenched power groups. There were some riots in Lancashire in 1779 and 1792, and a Manchester firm that pioneered a powerloom was burnt down. Yet cotton was unstoppable and must have seemed that way to contemporaries. Wool, however, was initially far larger and had an ancient tradition of professional organization and regulation. Laborers in the wool trades tried to use the political establishment for the purposes of stopping the new machines. In 1776 workers petitioned the House of Commons to suppress the jennies that threatened the livelihood of the industrious poor, as they put it. After 1789, Parliament passed sets of repressive laws (most famously the Combination Act of 1799), which in Horn's (2002) view were not only intended to save the regime from French-inspired revolutionary turmoil, but also to protect the Industrial Revolution from resistance "from below." Time and again, groups and lobbies turned to Parliament requesting the enforcement of old regulations or the introduction of new legislation that would hinder the machinery. Parliament refused. The old laws regulating the employment practices in the woollen industry were repealed in 1809, and the 250 year old Statute of Artificers was repealed in 1814. Lacking political support in London, the woolworkers tried extralegal means. As Randall has shown, in the West of England the new machines were met in most places by violent crowds, protesting against jennies, flying shuttles, gig mills, and scribbling machines (Randall, 1986; 1989). Moreover, in these areas magistrates were persuaded by fear or propaganda that the machine breakers were in the right. The tradition of violence in the West of England,

[76] Some of the following is based on Mokyr (1994).

writes Randall, deterred all but the most determined innovators. Worker resistance was responsible for the slow growth and depression of the industry rather than the reverse (Randall, 1989). The West of England, as a result, lost its supremacy to Yorkshire. Resistance in Yorkshire was not negligible either, but it was unable to stop mechanization. Violent protests, such as the Luddite riots, were forcefully suppressed by soldiers. As Paul Mantoux put it well many years ago, "Whether [the] resistance was instinctive or considered, peaceful or violent, it obviously had no chance of success" (Mantoux, 1928, p. 408).Had that not been the case, sustained progress in Britainwould have been severely hampered and possibly brought to an end.[77]

In other industries, too, resistance appeared, sometimes from unexpected corners. When Samuel Clegg and Frederick Windsor proposed a central gas distribution plan for London, they were attacked by a coalition that included the eminent scientist Humphry Davy, the novelist Walter Scott, the cartoonist George Cruickshank, insurance companies, and the aging James Watt (Stern, 1937). The steam engine was resisted in urban areas by fear of "smoky nuisances," and resistance to railroads was rampant in the first years of their incipience. Mechanical sawmills, widely used on the Continent, were virtually absent from Britain until the nineteenth century.[78] Even in medical technology, where the social benefits were most widely diffused, the status quo tried to resist. When Edward Jenner applied to the Royal Society to present his findings, he was told "not to risk his reputation by presenting to this learned body anything which appeared so much at variance with established knowledge and withal so incredible" (Keele, 1961,

---

[77] As Randall has shown, in the West of England the new machines were met by violent crowds, protesting against jennies, flying shuttles, gig mills, and scribbling machines (Randall, 1986; 1989). Moreover, in these areas magistrates were persuaded by fear or propaganda that the machine breakers were in the right. The tradition of violence in the West of England, writes Randall, deterred all but the most determined innovators. Worker resistance was responsible for the slow growth and depression of the industry rather than the reverse (Randall, 1989). The West of England, as a result, lost its supremacy in the wool industry to Yorkshire.

[78] The resistance against sawmills is a good example of attempts to use both legal and illegal means. It was widely believed in the eighteenth century that sawmills, like gigmills, were illegal although there is no evidence to demonstrate this. When a wind-powered sawmill was constructed at Limehouse (on the Thames, near London) in 1768, it was damaged by a mob of sawyers "on the pretence that it deprived many workmen of employment" (Cooney, 1991).

p. 94).[79] In medical technology, in general, resistance tended to be particularly fierce because many of the breakthroughs after 1750 were inconsistent with accepted doctrine, and rendered everything that medical professionals had laboriously learned null and void. It also tended, more than most other techniques, to incur the wrath of ethical purists who felt that some techniques in some way contradicted religious principles, not unlike the resistance to cloning and stem-cell research in our own time. Even such a seemingly enormously beneficial and harmless invention as anesthesia was objected on a host of philosophical grounds (Youngson, 1979, pp. 95-105; 190-98).

The two most famous cases of technology-related rioting in Britain are the Luddite riots between 1811 and 1816, and the Captain Swing riots of 1830-32. In both cases the riots were partially caused by technological innovation. To be sure, in Nottingham, where the Luddite troubles started, to be sure, there had been no technological change in the stocking frames and the anger of workmen was directed against low wages, work practices and similar issues. When the riots spread to Yorkshire, however, the finishers ("croppers") in the wool trade were directly motivated by the introduction of gig mills, shearing machines, and other machinery used in the finishing trades. The Yorkshire croppers were well-organized, and their main organization, "the Institution," was small and highly effective in organizing its members (Thomis, 1972 pp. 48-57). Their abortive attack on an advanced and mechanized mill at Rawfolds has become famous in the literature through its depiction in Charlotte Brontë's *Shirley* (Thomis, 1972; Thompson, 1963, pp. 559-65). In Lancashire, on the other hand, machine breaking during the Luddite riots occurred largely because they were a convenient target, not because of any deeply-felt anti-technological feeling. The history of machine breaking and violence against innovators is of course

---

[79] Jenner's famous discovery of the smallpox vaccine ran into the opposition of the inoculators concerned about losing their lucrative trade (Hopkins, 1983, p. 83). The source of the vaccine, infected animals, was a novelty and led to resistance in and of itself: Clergy objected to the technique because of the "iniquity of transferring disease from the beasts of the field to Man" (Cartwright, 1977, p. 86). Cartoonists depicted people acquiring bovine traits, and one woman complained that after he daughter was vaccinated she coughed like a cow and grew hairy (Hopkins, 1983, p. 84). Despite all this, of course, the smallpox vaccine was one of the most successful macroinventions of the period of the Industrial Revolution and its inventor became an international celebrity.

a complex story and not all cases of rioting were necessarily a response to technological change (Bohstedt, 1983, pp. 210-221). Moreover, machine breaking and rioting was just one of the ways in which resistance to technological change could manifest itself.

The Captain Swing riots were aimed, as is well-known, against the steam threshers. They bore in some ways a resemblance to the Luddite riots a decade and a half earlier, in that the resentment against machinery was aggravated by short-run fluctuations in the economy, and that the anger against new machinery was compounded by other grievances. The Swing riots were at least in part aimed against Irish migrant workers (Stevenson, 1979, p. 243). Yet the Captain Swing riots stand out because they were the only antitechnological movement in Britain, legal or extralegal, that was successful in slowing down the adoption of the technology altogether. The steam thresher against which they were aimed vanished from the South of England until the 1850s. As Hobsbawm and Rudé point out, the resistance against the machines was shared by some farmers and gentry. It was the first major successful act of "Luddism" in Britain in the nineteenth century, and it is perhaps symbolic that it occurred in the year typically (and arbitrarily) designed as the last year of the period known as the Industrial Revolution (Hobsbawm and Rudé, 1973, pp. 256-59, 317-23).

With the rise in the factory and the strengthening of the bargaining power of capitalists, authority and discipline might have reduced, at least for a while, the ability of labor to resist technological progress. The factory, however, did not solve the problem of resistance altogether; unions eventually tried to undermine the ability of the capitalist to exploit the most advanced techniques. Collective action by workers imposed an effective limit on the "authority" exercised by capitalists. Workers' associations tried to ban some new techniques altogether or tried to appropriate the entire productivity gains in terms of higher piece wages, thus weakening the incentive to innovate. On the other hand, such strikes often led to technological advances aimed specifically at crippling strikes (Bruland, 1982; Rosenberg, 1976, pp. 118-119).[80]

---

[80] The most famous example of an invention triggered by a strike was that of the self-acting mule, invented in 1825 by Richard Roberts at the prompting of Manchester manufacturers plagued by a strike of mule operators.

Conclusions: Technology, Growth, and the Rise of the Occident

In economic history, more so perhaps than in other disciplines, everything is a matter of degree, and there are no absolutes. The arguments made in this survey represent an interpretation that is by no means generally accepted. Many scholars have argued eloquently and persuasively for continuity rather than the view that something radical and deep changed in western society between 1760 and 1830. Almost every element we associate with the Industrial Revolution can be seen to have precedent and precursor. Some of these are quite valid (episodes of growth and "modernity" can be found in earlier periods, the use of coal and non-animate energy was expanding already in the centuries before the Industrial Revolution; agricultural productivity may have been as high in 1290 as it was in 1700; factory-like settings can be found in earlier periods). Others are based on misapprehensions (the aeolipiles built by Hero of Alexandria were *not* atmospheric steam engines). In the end, the debate on continuity can only be settled if we accept a criterion by which to judge the degree of continuity. If the criterion is economic growth, the continuity faction in the end will have to concede defeat, even if the loss is one in overtime. The Industrial Revolution *itself* was not a period of rapid economic growth, but it is clear beyond question that it set into motion an economic process that by the middle of the nineteenth century created a material world that followed a dynamic not hitherto experienced.

Not only that growth was faster and more geographically dispersed (covering by 1914 most of Europe, North America, other European offshoots, and Japan) than had been experienced by any economy before, it was sustainable. Unlike previous episodes, it kept rolling through the twentieth century. A moment of reflection will underline the enormity of this achievement. The twentieth century was in many ways a very bad century for the Western world: two horrid World Wars, a hugely costly depression, the collapse of international trade after 1914, the disastrous collectivist experiment in Russia extended to all of Eastern Europe in 1945, and the loss of its Colonial Empire — all of these should have pointed to catastrophe, misery, and a return to economic barbarism for the *Abendland*. Something similar may have happened in the fourteenth century, the disasters of which in some views set Europe's economy back for a century or more.

Yet by the 1990s, the gap between rich and poor nations is bigger than ever and Danny Quah's "twin peaks" are getting further and further

apart. Despite the huge setbacks, the engine that drove the Occident express had become so immensely powerful that it easily overwhelmed the twentieth century roadblocks that bad luck and human stupidity placed on its tracks. The Great Divergence train stormed on, undaunted.

Social scientists and historians discussing this issue are often accused of "triumphalism" which is paired with "Eurocentricity" or "Western-centricity." Whether the scholars who make such accusations actually mean to argue that gap in income and living standards is imaginary (or ephemeral), or whether they just feel that it is unjust and unfair, is sometimes hard to tell.[81] Yet if the rest of the world is to eventually enjoy the material comforts available to most people in the west or not, we should not give up on our attempt to understand "how the West did it."

To make some headway, if we want to understand why the West did what it did we should ask questions about the *when*. The consensus is that by 1750, the gap between the twin peaks was much smaller than it was today. If Europe was richer than the rest of the world, it was so by a margin that looks thin compared to what it is today. The so-called "California School" has been arguing indeed that living standards and measurable indicators of economic performance between China and Europe were not all that different by 1750.[82] If this is accepted, and if we are willing to take the Yang-Zhi delta as indicative of economic conditions of the non-European world, the current gap between rich and poor is largely the result of the Industrial Revolution and the events that followed it. Be that as it may, underneath its surface the European soil in 1500 already contained the seeds of the future divergence in 1750. There was, however, nothing inexorable about what happened after: the seeds need not have sprouted, they could have been washed away by the flood of wars, pulled out by rapacious tax collectors, or the young sprouts of future growth might have been burned by intolerant religious authorities. There could have been a Great *Convergence* after 1800 instead of what actually took place, in which Europe would have reverted back to the kind of economic performance prevalent in 1500. In the end, the economic history of technology — like all evolutionary sequences

---

[81] Such confusions mark especially the literature associated with Gunder Frank (1998) and Blaut (1993).

[82] See especially Wong (1997); Pomeranz (2000); Goldstone (2002).

— contains a deep and irreducible element of contingency. Not all that was had to be.

The question of "when" is important because it makes geographical explanations that explain Europe's success by its milder climate or conveniently located coal reserves less powerful, because these differences are time-invariant. Something had changed in Europe before the Industrial Revolution that destabilized the economic dynamic in the West, but not elsewhere. The question of "where" is also important. Britain was not "Europe" and even today there are some European regions that clearly are not part of the Western economic development pattern or very recent arrivals. On the other hand, a number of non-European nations have been able to join the "convergence club."

There are two alternative scenarios of the emergence of the gap. One is that, regardless of living standards and income in 1750, Europe was already deeply different in 1750 in many respects. In their different ways, David Landes (1998), Eric Jones (1981, 1988), Avner Greif (2003),and Angus Maddison (1998) subscribe to this view. By 1750 Europe had already had Calvin and Newton, Spinoza and Galileo, Bacon and Descartes. It had a commercial capitalism thriving especially in Atlantic Ports, a well-functioning monetary system, and the ability of rulers to tax their subjects had been constrained in complex but comparatively effective ways. It had universities, representative parliamentary bodies, embryonic financial institutions, powerful navies and armies, microscopes and printing presses. Its agriculture was gradually switching to more productive rotations, adopting new crops, and experimenting with animal breeding. Its manufacturing system was market-oriented and competitive. It had established the beginning of a public health system that had conquered the plague (still rampant elsewhere) and was making inroads against smallpox. Its ships, aided by sophisticated navigational instruments and maps, had subjugated and colonized already some parts of the non-European world and neither the Mongols nor the Ottoman Turks were a threat anymore. It drank tea, ate sugar, smoked tobacco, wore silk and cotton, and ate from better plates in coal- or peat heated homes. Its income per capita, as well as we can measure it, may have been little different from what it had been in the late middle ages (though Adam Smith disagreed), yet it was already ahead.

The alternative school emphasizes that many of these European features could be found in other societies, especially in China and Japan, and that when Europe and the Orient differed, the difference was not always

necessarily conducive to economic growth. Ch'ing China may not have been an open economy, but it had law and order, internal, a meritocratic bureaucracy peace and a great deal of medium- and long-distance trade within its borders. We need to be wary from the logical fallacy that all initial differences between Europe and China contributed to the outcome. Some of the initial difference may have actually worked the other way, so that the Great Divergence took place despite them. Others were ambiguous in their effect.[83] In order to understand what triggered Europe's economic miracle, we need to identify one more event that happened before the Industrial Revolution, happened in the right areas, and which can be connected logically to subsequent growth.

I have referred to this event as "the Industrial Enlightenment" and have attempted to show how it affected the two central elements of the Industrial Revolution, technology and institutions, and how these two elements then affected one another. The concept of Enlightenment I employ is somewhat different than what is customary. Not everything that is normally included in the historians' idea of the Enlightenment mattered, and not everything that mattered could be attributed to the Industrial Enlightenment. The emphasis on the Enlightenment illustrates how economists should think about culture and cultural beliefs as discussed in great length by Greif (2003). Culture mattered to economic development — how could it not? But we have to show the exact ways in which it mattered and through which channels it operated. I have argued that cultural beliefs changed in the eighteenth century, but beyond Greif's notion of cultural beliefs, I would include the metaphysical beliefs that people held about their environment and nature, and their attitudes toward the relationship between production and useful knowledge. It should also include their cultural beliefs about the possibility and desirability of progress and their notions of economic freedom, property, and novelty.

In that sense, at least, the Enlightenment may have been the missing link that economic historians have hitherto missed. Greif (XIII-17) points out that many of the institutional elements of modern Europe were already in

---

[83] An example is the European States System, often hailed as the element of competition which constrained and disciplined European governments into a more rational behavior, lest they weaken their military power. Yet the costs of wars may well have exceeded the gains, and the mercantilist policies that the States System triggered in the seventeenth century had doubtful effects on economic performance.

place in the late Middle Ages: individualism, man-made formal law, corporatism, self governance, and rules that were determined through an institutionalized process in which those are subject to them can be heard and have an input. Yet these elements did not trigger modern growth at that time, and it bears reflecting why not. The technological constraints were too confining, and the negative feedbacks too strong. The Baconian belief that nature is logical and understandable, that the understanding of nature leads to its control, and that control of nature is the surest route to increased wealth, was the background of a movement that, although it affected but a minute percentage of Europe's population, played a pivotal role in the emergence of modern growth. If culture can be said to matter, it did so because the prevailing ideology of knowledge among those who mattered started to change in a way it did not elsewhere. The eighteenth century Enlightenment, moreover, brought back many of the institutional elements of an orderly and civil society, together with the growing realization, most eloquently expressed by Adam Smith, that economic activity was not a zero-sum game and that redistributive institutions and rent-seeking are costly to society.

All the same, ideological changes and cultural developments are not the entire story. A desire for improvement and even the "right" kind of institutions by themselves do not produce *sustained* growth unless society produces new useful knowledge and unless the growth of knowledge can be sustained over time. Useful knowledge grows because in each society there are people who are creative and original, and motivated by some combination of greed, ambition, curiosity, and altruism. All four of those motives can be seen to be operating among the people who helped make the Industrial Revolution, often in the same people. Yet in order to be translated from personal predilections to facts on the ground and from there to economic growth, an environment that produced the correct incentives and the proper access to knowledge had to be there. The uniqueness of the European Enlightenment was that it created that kind of environment.

I have argued here that the experience of the past two centuries support the view that the production and utilization of knowledge are consistent with an interpretation that useful knowledge went through a phase transition in which it entered a critical region in which equilibrium concepts may no longer apply. This means that as far as future technological progress and economic growth are concerned, not even the sky is the limit. Science Fiction writers have known this all along.

References

Aghion, Philippe, and Peter W. Howitt. 1997. *Endogenous Growth Theory*. Cambridge, Mass.: MIT Press.

Allen, Robert C., and Cormac Ó Gráda. 1988. "On the Road Again with Arthur Young: English, Irish, and French Agriculture during the Industrial Revolution." *Journal of Economic History* 48, no. 1 (March), pp. 93–116.

Antrás, Pol and Voth, Joachim. 2003. Factor Prices and Productivity Growth during the British Industrial Revolution." *Explorations in Economic History*, Vol. 40, pp. 52-77.

Baumol, William J. 2002. *The Free-Market Innovation Machine: Analyzing the Growth Miracle of Capitalism*. Princeton, N.J.: Princeton University Press.

Bohstedt, John. 1983. *Riots and Community Politics in England and Wales, 1790–1810*. Cambridge, Mass.: Harvard University Press.

Botticini, Maristella and Eckstein, Zvi. "From Farmers to Merchants: a Human Capital Interpretation of Jewish Economic history." Unpublished ms., Boston University, Jan. 2003.

Britnell, Richard H. 1996. *The Commercialization of English Society*, *1000-1500*. Manchester: Manchester University Press.

Brock, William H. 1992. *The Norton History of Chemistry*. New York: W. W. Norton.

Brown, G. I. 1999. *Scientist, Soldier, Statesman, Spy: Count Rumford*. Gloucestershire, Eng.: Sutton Publishing.

Bruland, Tine. 1982. "Industrial Conflict as a Source of Technical Innovation: Three Cases." *Economy and Society* 11:91-121.

Buchanan, R. A. 1985. "The Rise of Scientific Engineering in Britain." *British Journal for the History of Science* 18, no. 59 (July), pp. 218–33.

Burton, Anthony. 2000. *Richard Trevithick: Giant of Steam*. London: Aurum Press.

Cain, Louis, and Elyce Rotella. 2001. "Death and Spending: Did Urban Mortality Shocks Lead to Municipal Expenditure Increases?" *Annales de Démographie Historique* 1, pp. 139-54.

Cardwell, Donald S. L. 1971. *From Watt to Clausius: The Rise of Thermodynamics in the Early Industrial Age.* Ithaca, N.Y.: Cornell University Press.

———. 1972. *Turning Points in Western Technology*. New York: Neale Watson, Science History Publications.

Cartwright, F. F. 1977. *A Social History of Medicine*. London: Longman.

Cervellati, Matteo and Sunde, Uwe. 2002. "Human Capital Formation, Life Expectancy, and the Process of Economic Development," IZA (Bonn) discussion papers 585, Sept.

Chitnis, Anand. 1976. *The Scottish Enlightenment.* London: Croom Helm.

Cohen, Jack, and Ian Stewart. 1994. *The Collapse of Chaos: Discovering Simplicity in a Complex World*. Harmondsworth, Eng.: Penguin.

Cole, Arthur H., and George B. Watts, 1952. *The Handicrafts of France as Recorded in the Descriptions des Arts et Métiers 1761–1788*. Boston: Baker Library.

Constant, Edward W. 1980. *The Origins of the Turbojet Revolution.* Baltimore: Johns Hopkins Press.

Cooney, E. W. 1991. "Eighteenth Century Britain's Missing Sawmills: A Blessing in Disguise?" *Construction History* 7, pp. 29–46.

Cowan, Robin, and Dominique Foray. 1997. "The Economics of Codification and the Diffusion of Knowledge." *Industrial and Corporate Change* 6, no. 3 (Sept.), pp. 595–622.

Crouch, Tom. 1989. *The Bishop's Boys: A Life of Wilbur and Orville Wright*. New York: W. W. Norton.

Darnton, Robert. 1979. *The Business of Enlightenment* (Cambridge, MA: Harvard University Press).

David, Paul A. 1997. "Reputation and Agency in the Historical Emergence of the Institutions of 'Open' Science." Unpublished ms., Oxford University.

———.1998."The Collective Cognitive Performance of 'Invisible Colleges'." Presented to the Santa Fe Institute Workshop "The Evolution of Science."

Davids, Karel. 2000. "Patents and Patentees in the Dutch Republic between c. 1580 and 1720." *History and Technology* 16, pp. 263–83.

———. 2001. "Windmills and the Openness of Knowledge: Technological Innovation in a Dutch Industrial District, the Zaanstreek, c. 1600–1800." Unpublished paper, presented to the Annual Meeting of the Society for the History of Technology, San Jose, Calif.

De Vries, Jan and Van Der Woude, A.M.. 1997. *The First Modern Economy: Success, Failure, and Perseverance of the Dutch Economy, 1500-1815.* Cambridge: Cambridge University Press.

Dobbs, B.J.T. 1990. "From the Secrecy of Alchemy to the Openness of Chemistry." In Tore Frängsmyr, ed., *Solomon's House Revisited: The Organization and Institutionalization of Science,* pp. 75–94. Canton, Mass.: Science History Publishing.

Donovan, A. L. 1975. *Philosophical Chemistry in the Scottish Enlightenment*. Edinburgh: at the University Press.

Eamon, William. 1990. "From the Secrets of Nature to Public Knowledge" in David C. Lindberg and Robert S. Westman, eds., *Reappraisals of the Scientific Revolution,* pp. 333–65. Cambridge: Cambridge University Press.

———. 1994. *Science and the Secrets of Nature*. Princeton, N.J.: Princeton University Press.

Easterlin, Richard. 1981. "Why isn't the Whole World Developed?" *Journal of Economic History* Vol. 41, No. 1, pp. 1-19.

Eisenstein, Elizabeth. 1979. *The Printing Press as an Agent of Change.* Cambridge: Cambridge University Press.

Ekelund, Robert B. Jr., and Tollison, Robert D., 1981. *Mercantilism as a Rent-Seeking Society*. College Station: Texas A&M University Press.

Ekelund, Robert B. Jr., and Tollison, Robert D., 1997. *Politicized Economies: Monarchy, Monopoly, and Mercantilism.* College Station: Texas A&M University Press.

Engerman, Stanley S. 1994. "The Industrial Revolution Revisited." In Graeme Donald Snooks, ed., *Was the Industrial Revolution Necessary?* London: Routledge.

Fairchilds, Cissie. 1992. "A Comparison of the 'Consumer Revolutions' in Eighteenth-Century England and France." Unpublished paper, submitted to the Economic History Association Annual Meeting, Boston.

Fox, Robert. 1998. "Science, Practice and Innovation in the Age of Natural Dyes, 1750–1860." In Maxine Berg and Kristin Bruland, eds., *Technological Revolutions in Europe,* pp. 86–95. Cheltenham, Eng.: Edward Elgar,

Fox, Robert, and Anna Guagnini. 1999. *Laboratories, Workshops, and Sites: Concepts and Practices of Research in Industrial Europe, 1800–1914*. University of California, Berkeley, Office for History of Science and Technology.

Galor, Oded and Weil, David .2000. "Population, Technology, and Growth." *American Economic Review* 90 no. 4 (Sept.), pp. 806-828.

Galor, Oded and Moav, Omer. 2002. "Natural Selection and the Origins of Economic Growth." *Quarterly Journal of Economics*, Vol. 117, No. 4 (Nov.), pp. 1133-91.

Gillispie, Charles C. 1957. "The Natural History of Industry." *Isis* 48, pp. 398–407.

Goddard, Nicholas. 1989, "Agricultural Literature and Societies," in G.E. Mingay, ed., *The Agrarian History of England and Wales*, Vol. VI, 1750-1850. Cambridge: Cambridge University Press, pp. 361-83.

Golinski, Jan. 1988. "Utility and Audience in Eighteenth Century Chemistry: Case Studies of William Cullen and Joseph Priestley." *British Journal for the History of Science* 21, no. 68, pt. I (March), pp. 1–31.

———. 1992. *Science as Public Culture: Chemistry and Enlightenment in Britain, 1760–1820*. Cambridge: Cambridge University Press.

Greif, 2003 *Comparative and Historical Institutional Analysis: a Game-Theoretical Perspective*. Forthcoming, Cambridge University Press.

Griffiths, John. 1992. *The Third Man: the Life and Times of William Murdoch, inventor of gaslight*. London: André Deutsch.

Haber, L.F. 1958. *The Chemical Industry during the Nineteenth Century*. Oxford: At the Clarendon Press.

Hall, A. Rupert. 1974. "What Did the Industrial Revolution in Britain Owe to Science?" In Neil McKendrick, ed., *Historical Perspectives: Studies in English Thought and Society*. London: Europa Publications.

Herman, Arthur. 2001 *How the Scots Invented the Modern World*. New York: Crown.
Hilaire-Pérez, Liliane. 2000. *L'invention technique au siècle des lumières*. Paris: Albin Michel.
Hobsbawm, Eric J., and George Rudé. 1973. *Captain Swing*. Harmondsworth, Eng.: Penguin Books.
Hopkins, Donald R. 1983. *Princes and Peasants: Smallpox in History*. Chicago: University of Chicago Press.
Horn, Jeff. 2003. "Machine Breaking in England and France during the Age of Revolution," unpub. ms., Manhattan College.
Hucker, Charles O. 1975. *China's imperial past: an introduction to Chinese history and culture*. Stanford: Stanford University Press.
Hudson, Derek, and Kenneth W. Luckhurst. 1954. *The Royal Society of Arts, 1754–1954.* London: John Murray.
Inkster, Ian. 1991. *Swcience and Technology in History: an Approach to Industrial Development.* New Brunswick, NJ: Rutgers University Press.
James, Harold. 1990. "The German Experience and the Myth of British Cultural Exceptionalism." In Bruce Collins and Keith Robbins, eds.*, British Culture and Economic Decline*, pp. 91–128. New York: St. Martin's Press.
Jones, Eric L. 1981. The *European Miracle: Environments, Economies and Geopolitics in the History of Europe and Asia. 2d ed.* 1987. Cambridge: Cambridge University Press.
———. 1988. *Growth Recurring*. Oxford: Oxford University Press.
Jones, Charles I. 2001. "Was an Industrial Revolution Inevitable? Economic Growth over the very long run." *Advances in Macroeconomics,* (Berkeley Electronic Press) Vol. 1, No. 2, pp. 1-43.
Kauffman, Stuart A. 1995. *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*. New York: Oxford University Press.
Keele, K. D. 1961. "The Influence of Clinical Research on the Evolution of Medical Practice in Britain." In F.N.L. Poynter, ed., *The Evolution of Medical Practice in Britain,* pp. 81–96. London: Pitman Medical Publishing
Kelly, Fred C. 1943. *The Wright Brothers.* NewYork: Harcourt, Brace & Co.
Keyser, Barbara Whitney. 1990. "Between Science and Craft: the Case of Berthollet and Dyeing." *Annals of Science* 47, no. 3 (May), pp. 213–60.
Khan, B. Zorina. 2002. *'The Fuel of Interest': Patents and Copyrights in American Economic Development.* Cambridge: Cambridge University Press.
Khan, B. Zorina, and Kenneth L. Sokoloff. 1993. "'Schemes of Practical Utility': Entrepreneurship and Innovation among 'Great Inventors' in the United States, 1790–1865." *Journal of Economic History* 53, no. 2 (June), pp. 289–307.
———. 1998. "Patent Institutions, Industrial Organization, and Early Technological Change: Britain and the United States, 1790–1850." In Maxine

Berg and Kristin Bruland, eds., *Technological Revolutions in Europe,* pp. 292–313. Cheltenham, Eng.: Edward Elgar.

———. 2001. "The Early Development of Intellectual Property Institutions in the United States." *Journal of Economic Perspectives* 15, no. 2 (Spring), pp. 1–15.

Kirby, Peter. 2003. *Child Labor in Britain, 1750-1870* . Basingstoke: Palgrave-MacMillan.

Knowles, K. and Robertson, D. 1951. "Differences between the wages of skilled and unskilled workers, 1880-1950." *Bulletin of the Oxford University Institute of Statistics*, (April), pp. 109-27.

Kranakis, Eda. 1992. "Hybrid Careers and the Interaction of Science and Technology." In Peter Kroes and Martijn Bakker, eds., *Technological Development and Science in the Industrial Age,* pp. 177–204. Dordrecht: Kluwer.

Kremer, Michael. 1993. "Population Growth and Technological change: One million BC to 1990." *Quarterly Journal of Economics* Vol. 108, No. 4 (August), pp. 681-716.

Kronick, David A. 1962. *A History of Scientific and Technical Periodicals*. New York: Scarecrow Press.

Landes, David S. 1998. *The Wealth and Poverty of Nations: Why Some Are So Rich and Some So Poor*. New York: W. W. Norton.

Lerner, Josh. 2000. "150 years of Patent Protection." Working paper 7477. National Bureau of Economic Research, Cambridge, Mass.

Loasby, Brian J. 1999. *Knowledge, Institutions, and Evolution in Economics*. London: Routledge.

Lowood, Henry. 1991. *Patriotism, Profit, and the Promotion of Science in the German Enlightenment: the economic and scientific societies, 1760-1815*. New York: Garland Pub.

Lucas, Robert E. 2002. *Lectures on Economic Growth*. Cambridge, Mass.: Harvard University Press.

Macfarlane, Alan, and Gerry Martin. 2002. *Glass: A World History*. Chicago: University of Chicago Press.

MacLeod, Christine. 1988. *Inventing the Industrial Revolution: The English Patent System, 1660–1880.* Cambridge: Cambridge University Press.

Maddison, Angus. 1998. *Chinese Economic Performance in the Long-run*. Paris: OECD.

Mantoux, Paul. [1928] 1961. *The Industrial Revolution in the Eighteenth Century*. New York: Harper Torchbooks.

McClellan, James E. III. 1981. "The Academie Royale des Sciences, 1699-1793: A Statistical Portrait." *Isis* Vol. 72, No. 4. pp. 541-567.

———. 1985. *Science Reorganized: Scientific Societies in the Eighteenth century.* New York: Columbia University Press.

McCloy, Shelby T. 1952. *French Inventions of the Eighteenth Century*. Lexington: University of Kentucky Press.
McKendrick, Neil. 1973. "The Role of Science in the Industrial Revolution." In Mikuláš Teich and Robert Young, eds., *Changing Perspectives in the History of Science*. London: Heinemann.
Merton, Robert K. [1938] 1970. *Science, Technology, and Society in Seventeenth century England.* 2d ed. New York: Fertig.
Mitch, David. 1998. "The Role of Education and Skill in the British Industrial Revolution." In Joel Mokyr, ed., *The British Industrial Revolution: An Economic Perspective*, 2d ed., pp. 241–79. Boulder, Colo.: Westview Press.
Mokyr, Joel. 1990. *The Lever of Riches: Technological Creativity and Economic Progress*. New York: Oxford University Press.
———. 1994. "Progress and Inertia in Technological Change." In John James and Mark Thomas, eds., *Capitalism in Context: Essays in Honor of R. M. Hartwell*, pp. 230–54. Chicago: University of Chicago Press.
———. 1998a. "The Political Economy of Technological Change: Resistance and Innovation in Economic history." In Maxine Berg and Kristine Bruland, eds., *Technological Revolutions in Europe*, pp. 39–64. Cheltenham: Edward Elgar.
———. 1998b. "Editor's Introduction: The New Economic history and the Industrial Revolution." In Joel Mokyr, ed., The British Industrial Revolution: An Economic Perspective. Boulder: Westview Press, pp. 1-127.
———. 2000. "The Industrial Revolution and the Netherlands: Why did it not happen?" *De Economist* (Amsterdam) 148, no. 4 (Oct.), pp. 503–20.
———. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton: Princeton University Press.
Mokyr, Joel. 2003a. "Useful Knowledge as an Evolving System: the view from Economic history," in Larry Blume and Steven Durlauf eds., *The Economy as and Evolving Complex System* Vol. III,, Oxford University Press, forthcoming.
———. 2003b. "Mercantilism, the Enlightenment, and the Industrial Revolution," Presented to the Conference in Honor of Eli F. Heckscher, Stockholm, (May).
———. 2003c. "Accounting for the Industrial Revolution," in Paul Johnson and Roderick Floud, eds., *The Cambridge Economic History of Britain, 1700-2000.* Forthcoming, Cambridge University Press.
Morus, Iwan Rhys. 1998. *Frankenstein's Children : Electricity, Exhibition, and Experiment in Early-Nineteenth-Century London*. Princeton, N.J.: Princeton University Press.
Murmann, Johann Peter. 1998. "Knowledge and Competitive Advantage in the Synthetic Dye Industry, 1850–1914." Ph. D. dissertation, Columbia University.
Musson, A. E., and Eric Robinson. 1969. *Science and Technology in the Industrial Revolution.* Manchester: Manchester University Press.

Nelson, Richard R. 1994. "Economic Growth through the Co-evolution of Technology and Institutions." In Loet Leydesdorff and Peter Van Den Besselaar, eds., *Evolutionary Economics and Chaos Theory: New Directions in Technology Studies.* New York: St. Martin's Press.

Nelson, Richard R., and Sidney Winter. 1982. *An Evolutionary Theory of Economic Change*. Cambridge, Mass.: The Belknap Press.

North, Douglass C. 1990. *Institutions, Institutional Change, and Economic Performance*. Cambridge: Cambridge University Press.

O'Brien, Patrick. 2003. "The Hanoverian State and the Defeat of the Continental State," Presented to the Conference in honor of Eli F. Heckscher, Stockholm, May.

Olsson, Ola. 2000. "Knowledge as a Set in Idea Space: an Epistemological View on Growth," *Journal of Economic Growth* 5(3), pp. 253-76.

———. 2003. "Technological Opportunity and Growth." unpub. Ms., Göteborg University.

Ormrod, David. 2003. *The Rise of Commercial Empires: England and the Netherlands in the Age of Mercantilism, 1650-1770*. Cambridge: Cambridge University Press.

Pannabecker, John R. 1996, "Diderot, Rousseau, and the Mechanical Arts: Disciplines, Systems, and Social Context." *Journal of Industrial Teacher Education* vol. 33, No. 4, pp. 6-22.

Picon, Antoine. 2001. "Technology." In Michel Delon, ed., *Encyclopedia of the Enlightenment*, pp. 1317–23. Chicago: Fitzroy Dearborn.

Polanyi, Michael. 1962. *Personal Knowledge: Towards a Post-Critical Philosophy*. Chicago: Chicago University Press.

Pomeranz, Kenneth. 2000. *The Great Divergence: China, Europe, and the Making of the Modern World Economy*. Princeton, N.J.: Princeton University Press.

Porter, Roy. 1990. *English Society in the 18th Century*. 2nd ed. London: Penguin Books.

———. 2000. *The Creation of the Modern World: The Untold Story of the British Enlightenment*. New York: W.W. Norton.

Price, Derek J. de Solla. 1984a. "Notes towards a Philosophy of the Science/Technology Interaction" In Rachel Laudan, ed., *The Nature of Knowledge: are Models of Scientific Change Relevant?* Dordrecht: Kluwer.

———. 1984b. "Of Sealing Wax and String." *Natural History*, no. 1, pp. 49–56.

Priestley, Joseph. 1768. *An Essay on the First Principles of Government and on the Nature of Political, Civil and Religious Liberty.* London: Printed for J. Doosley in Pall Mall.

Randall, Adrian J. 1986. "The Philosophy of Luddism: The Case of the West of England Workers, ca. 1790–1809." *Technology and Culture* 27, no. 1 (Jan.), pp. 1–17.

———. 1989. "Work, Culture and Resistance to Machinery in the West of England Woollen Industry." In Pat Hudson, ed., *Regions and Industries: A Perspective on the Industrial Revolution in Britain*, pp. 175–98. Cambridge: Cambridge University Press.

Reiter, Stanley. 1992. "Knowledge, Discovery and Growth." Discussion Paper 1011. Northwestern University Center for Mathematical Studies in Economics and Management Sciences.

Roche, Daniel. 1998. *France in the Enlightenment*. Harvard: Harvard University Press.

Root, Hilton. 1994. *The Fountain of Privilege: Political Foundations of Markets in Old Regime France and England*. Berkeley and Los Angeles: University of California Press.

Rosenberg, Nathan. 1976. *Perspectives on Technology*. Cambridge: Cambridge University Press.

———. 1994. *Exploring the Black Box.* New York: Cambridge University Press.

Sandberg Lars G. "The Case of the Impoverished Sophisticate: Human Capital and Swedish Economic Growth before World War I." *The Journal of Economic History*, Vol. 39, No. 1,(March), pp. 225-241.

Schiff, Eric. 1971. *Industrialization without National Patent.* Princeton, N.J.: Princeton University Press.

Schultz, Theodore W. 1975 "The Value of the Ability to Deal with Disequilibria." *Journal of Economic Literature,* Vol. 13, No. 3. (Sept), pp. 827-846.

Shapin, Steven. 1994. *The Social History of Truth*. Chicago: University of Chicago Press.

———. 1996. *The Scientific Revolution*. Chicago: University of Chicago Press.

Shleifer, Andrei and Vishny, Robert. 1998. *The grabbing hand : government pathologies and their cures*. Cambridge, MA: Harvard University Press.

Simon, Julian L. 1977. *The Economics of Population Growth*. Princeton: Princeton University Press.

———. 2000. *The Great Breakthrough and Its Cause*, edited by Timur Kuran. Ann Arbor: The University of Michigan Press.

Smith, Adam. [1776] 1976. *The Wealth of Nations*. Edited by Edwin Cannan. Chicago: University of Chicago Press.

Smith, John Graham. 1979. *The Origins and Early Development of the Heavy Chemical Industry in France.* Oxford: Clarendon Press.

———. 2001. "Science and Technology in the Early French Chemical Industry." Unpublished paper, presented to the colloquium on "Science, techniques, et Sociétés," Paris.

Snelders, H.A.M. 1992. "Professors, Amateurs, and Learned Societies." In Margaret Jacob and Wijnand W. Mijnhardt, eds., *The Dutch Republic in the Eighteenth century: Decline, Enlightenment, and Revolution*. Ithaca: Cornell University Press.

Snooks, 1994

Spadafora, David. 1990. *The Idea of Progress in Eighteenth-Century Britain.* New Haven: Yale University Press.

Sparrow, W. J. 1964. *Knight of the White Eagle*. London: Hutchinson and Co.

Stern, Bernhard J. 1937. "Resistances to the Adoption of Technological Innovations." In *Technological Trends and National Policy*. Washington, D.C.: United States Government Printing Office.

Stevenson, John. 1979. *Popular Disturbances in England, 1700–1870*. New York: Longman.

Thomis, Malcolm. 1972. *The Luddites*. New York: Schocken

Thompson, E. P. 1963. *The Making of the English Working Class*. New York: Vintage Books.

Thorne, Stuart. 1986. *The History of Food Preservation*. Totowa, N.J.: Barnes and Noble Books.

Tranter, N.L. 1985. *Population and Society, 1750-1940*. Burnt Mill: Longman.

Travis, Anthony. 1989. "Science as Receptor of Technology: Paul Ehrlich and the Synthetic Dyestuff Industry," *Science in Context* 3, no. 2 (Autumn), pp. 383–408.

Vickers, Brian. 1992. "Francis Bacon and the Progress of Knowledge." *Journal of the History of Ideas*, Vol. 53, No. 2 (July-Sept.), pp. 493-518.

Vincenti, Walter. 1990. *What Engineers Know and How They Know It.* Baltimore: Johns Hopkins University Press.

———. 2000. "Real-World Variation-Selection in the Evolution of Technological Form: Historical Examples." In John Ziman, ed., *Technological Innovation as an Evolutionary Process,* pp. 174–89. Cambridge: Cambridge University Press.

Weatherill, Lorna, 1988. *Consumer Behaviour and Material Culture in Britain, 1660-1760*. New York: Routledge.

Wilson, Charles. 1969. "Taxation and the Decline of Empires: an Unfashionable Theme" in *Economic History and the Historians,* pp. 117-27. London: Weidenfeld and Nicolson.

Wong, R. Bin 1997. *China Transformed: Historical Change and the Limits of European Experience. Ithaca: Cornell University Press.*

Wood, Henry Trueman. 1913. *A History of the Royal Society of Arts*. London: John Murray.

Wuketits, Franz. 1990. *Evolutionary Epistemology and Its Implications for Humankind*. Albany: SUNY Press.

Youngson, A. J. 1979. *The Scientific Revolution in Victorian Medicine.* New York: Holmes and Meier Publishers.

Zilsel, Edgar. 1942. "The Sociological Roots of Science." *American Journal of Sociology* 47, no. 4 (Jan.), pp. 544—60.

Ziman, John. 1976. *The Force of Knowledge*. Cambridge: Cambridge University Press.

Ziman, John , ed., *Technological Innovation as an Evolutionary Process*. Cambridge: Cambridge University Press, 2000.

General Purpose Technologies

Boyan Jovanovic and Peter L. Rousseau*

January 2003

Abstract

Electricity and IT are perhaps the two most important GPTs to date. We analyze how the U.S. economy reacted to them. The Electricity and IT eras are similar, but they also different in important ways. Electrification was more broadly adopted, whereas IT seems to be technologically more revolutionary. The productivity slowdown is stronger in the IT era but the ongoing spread of IT and its continuing precipitous price decline are reasons for optimism about growth in the coming decades.

1 Introduction

The term "general-purpose technology," or GPT, has seen extensive use in recent treatments of the role of technology in economic growth, and is usually reserved for changes that transform both household life and the ways in which firms conduct business. Steam, electricity, and information technology (IT) are often classified as GPTs for this reason. They affected the whole economy.

As David (1991) pointed out some years ago, a GPT does not deliver productivity gains immediately upon arrival. Figure 1 shows the evolution of the growth in output per man hour in the U.S. economy over the past 130 years, with periods of rapid diffusion of the two major GPTs shaded and the dashed line representing long-term trends as generated with the Hodrick-Prescott (H-P) filter.[1] Productivity growth was apparently quite rapid during the heyday of steam power (circa. 1870), but fell as electrification arrived in the 1890s, with the defining moment probably being the

*NYU and the University of Chicago, and Vanderbilt University. We thank the National Science Foundation for support, and J. Cummins, B. Hobijn, J. Lerner and G. Violante for providing us with data. This is a proposed chapter in the forthcoming Handbook of Economic Growth.

[1] Output per man-hour in the business, non-farm sector is from John Kendrick [U.S. Bureau of the Census (1975, Series D684, p. 162)] for 1889-1947, and from the Bureau of Labor Statistics (2002) for 1948-2001. For 1874-1889, we use Kendrick's decadal averages for 1869-79 and 1879-89, and assume a constant growth rate from 1874-84 and 1885-89.

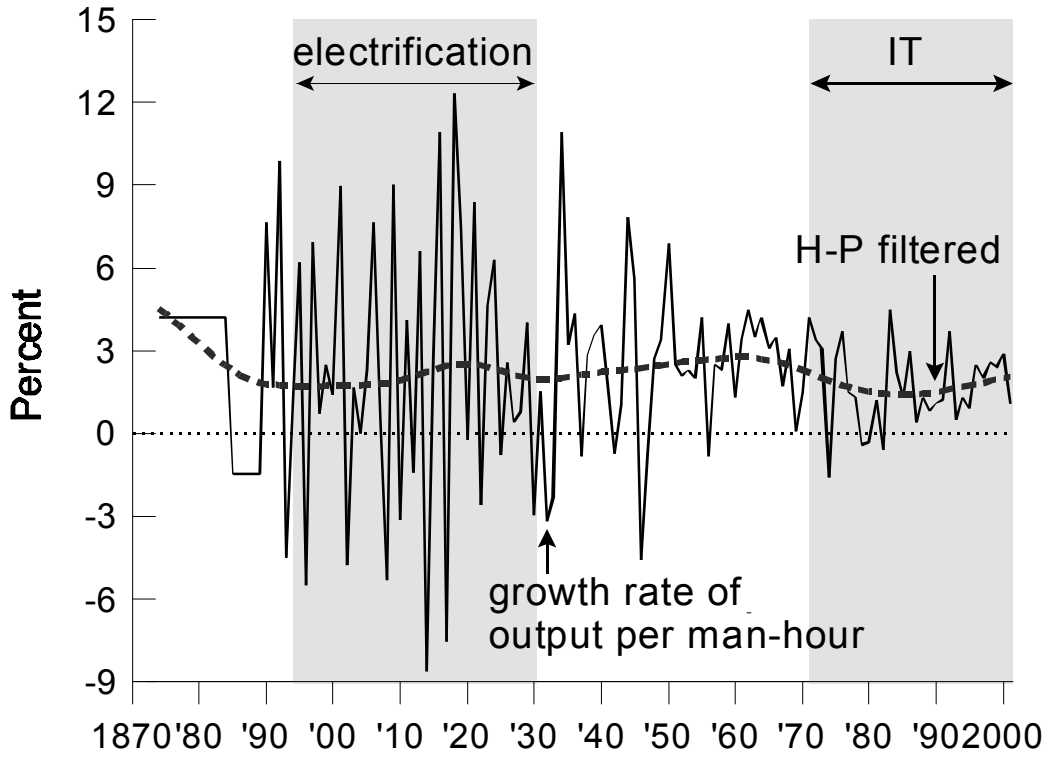


Figure 1: Annual growth in output per man-hour.

startup of the first hydro-electric facility at Niagara Falls in 1894. It was only in the period after 1915, which saw the diffusion of secondary motors and the widespread establishment of centralized power grids, that electricity finally pervaded businesses and households more generally and the productivity numbers began to rise. Figure 1 also shows that the arrival of IT, which we date with Intel's invention in 1971 of the "4004 computer chip" (the key component of the personal computer), did not reverse the decline in productivity growth that had begun more than a decade earlier. To some extent it seems that we are still waiting for computers to show up in the productivity figures.

But it is not obvious that the startup of the Niagara Falls dam in 1894 and Intel's invention of the 4004 chip in 1971 should define the birth of the two GPTs. Indeed, the reader may wonder how we choose the dating for the two GPT eras. In fact, the dates do coincide with periods of adoption: Net adoption of the two GPTs picks up about where the shading begins and, in the case of electrification, adoption reaches a plateau in 1929, whereas new adoption of IT is still rising today so that, on that criterion, the IT era still continues.

1.1 What is a GPT?

Each shaded area in Figure 1 contains a growth slowdown. Will the growth slowdown of the current IT era be followed by a rise in growth in the first half of the 21st century? If the second shaded area in Figure 1 is in some "fundamental" respects like the first shaded area, then we can expect growth to pick up over the next several decades. In Jovanovic and Rousseau (2002a) we have argued that the first half of the 21st century will have higher growth than, say, the 1950s and 1960s. Gordon (2000), on the other hand, is pessimistic, arguing that IT does not measure up to electricity and that it will not have such positive results. This chapter will conclude that the two eras are indeed similar.

So, what are these "fundamental" features of the two GPTs that we may attempt to compare? More generally, what criteria can one use to distinguish a GPT from other technologies? Bresnahan and Trajtenberg (1995) argue that a GPT should have the following three characteristics:

1. *Pervasiveness*: The GPT should spread to most sectors.

2. *Improvement*: The GPT should get better over time and, hence, should keep lowering the costs of its users.

3. *Innovation spawning*: The GPT should make it easier to invent and produce new products or processes.

Most technologies possess each of these characteristics to some degree, and therefore a GPT cannot differ qualitatively from these other technologies. Moreover, this is a short list which we shall later broaden. Nevertheless, we shall begin with measures of these three characteristics in the next section. But first, we summarize our findings:

1.2 Summary of findings

The evidence shows similarities and differences between the electrification and the IT eras. Electrification was more pervasive (#1), whereas IT has a clear lead in terms of improvement (#2) and innovation spawning (#3). Let us list the similarities and differences in more detail.

1.2.1 Similarities between the electrification and IT eras

1. In both of the GPT eras growth is below rates attained in the decades immediately preceding.

2. Measures of reallocation and invention – startups, exits, patents, trademarks, and investment by new firms relative to incumbents – are all higher during the GPT eras.

3. Private consumption rose gradually during each GPT era.

4. Real interest rates are about the same during the two GPT eras, and about three percentage points higher than in the middle 40 years of the 20th century.

1.2.2 Differences between the electrification and IT eras

1. Innovation measures are growing much faster for IT than for Electrification – patents and trademarks surge much more strongly during the IT era, and the price of IT is falling 100 times faster, at least, than did the price of electricity.

2. IT is spreading more slowly than did electrification, and its net adoption still continues to rise in the United States.

3. The productivity slowdown is stronger in the IT era.

4. No comparable sudden collapse of the stock markey occured early on in the Electrification era.

5. The Electrification era saw a surplus in the U.S. trade balance, in part surely because Europe had to finance a string of wars, whereas the IT era finds the United States in a trade deficit

The differences seem to outnumber the similarities. Yet the overall evidence clearly supports the view that technological progress is uneven, that it does entail the episodic arrival of GPTs, that these GPTs bring on turbulence and lower growth early on and higher growth and prosperity later.

2 Measuring the three characteristics of a GPT

As suggested in Figure 1, we shall choose electricity and IT as our candidate GPTs, and the measures that we provide will pertain mostly to these two technologies. In passing, we shall also touch upon steam and internal combustion. The three subsections below report, in turn, various measures of each characteristic – pervasiveness, improvement, and innovation – for the two GPTs at hand.

2.1 Pervasiveness of the GPT

The first characteristic is the technology's pervasiveness. We begin with the aggregates and then look in more detail at industrial sectors.

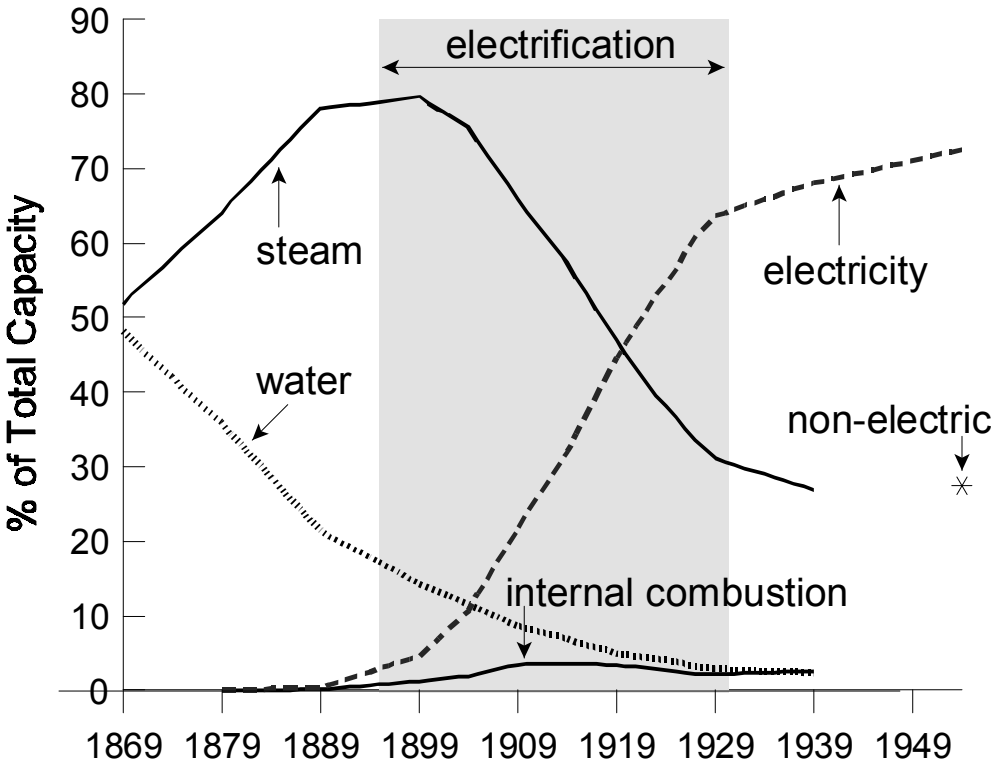


Figure 2: Shares of total horsepower generated by the main sources in U.S. manufacturing, 1869-1954.

2.1.1 Pervasiveness in the aggregate

We would like to track the evolution of GPTs using continuous time series from about 1850 to the present, but we do not have data that consistently cover the entire stretch of time, and thus will need to break this period into two overlapping segments: 1869-1954 and 1947-2001.

Figure 2 shows the shares of total horsepower in manufacturing by power source from 1869 to 1954.[2] The period covers the fall of water wheels and turbines, the rise and fall of steam engines and turbines, the rise and gradual flattening out of the internal combustion engine, and the sharp rise in the use of primary and secondary electric motors. The symmetry of the plot is striking in that, with the exception

[2]We construct the shares of total horsepower capacity in manufacturing as ratios of each power source (DuBoff ,1964, table 14, p. 59) to the total (table 13, p. 58). DuBoff estimates these quantities in 1869, 1879, 1889, 1899, 1904, 1909, 1914, 1919, 1923, 1925, 1927, 1929, 1939, and 1954, and we linearly interpolate between these years. This source does not include a breakdown of non-electrical capacity (i.e., water, internal combustion, and steam) after 1939, and so we mark the broader-defined "non-electrical" share for 1954 with an asterisk.

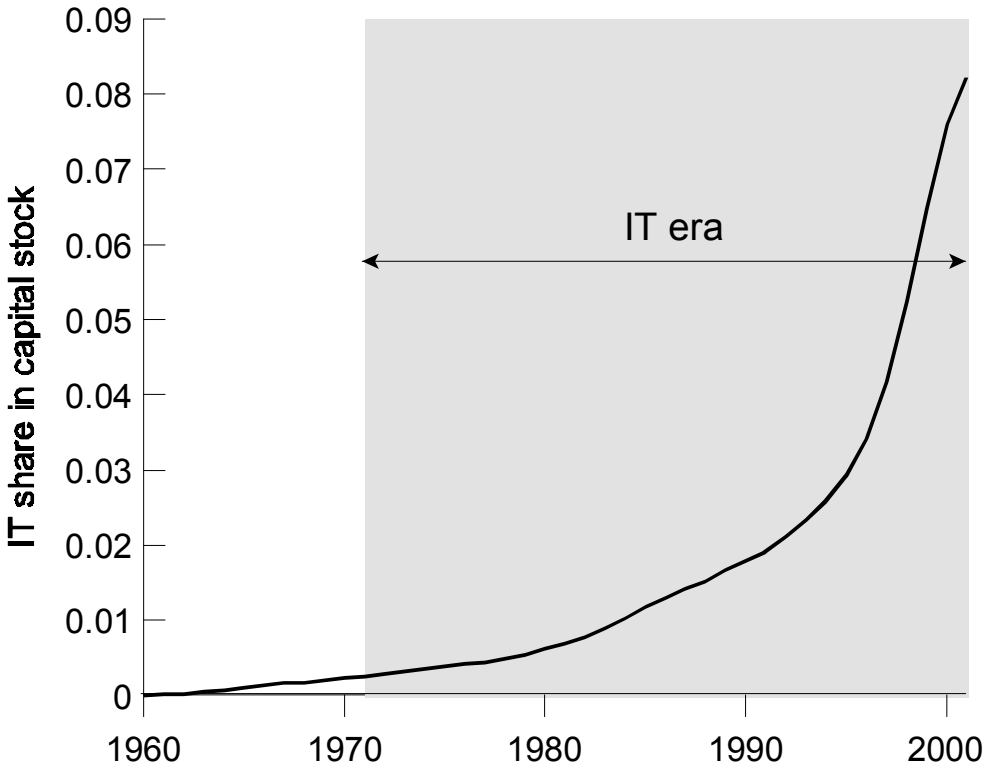


Figure 3: Shares of computer equipment and software in the aggregate capital stock, 1960-2001.

of internal combustion, power-generating technologies seem to have led for the most part sequential existences. The relative brevity of the entire steam cycle, which rises and falls within a period of 50-60 years, suggests that the technology that replaced it, Electricity, was important enough to force change quickly among manufacturers. In contrast, the decline of water power was more gradual. Moreover, if we could continue the graph to the present, electricity would surely still command a very high share of manufacturing power as a new source (e.g., solar power?) has not yet emerged to replace it. The persistence of electricity as the primary power source, even though its diffusion throughout the manufacturing sector was complete decades ago, helps to identify it as one of the breakthrough technologies of the modern era.

Figure 3 shows the diffusion of computers in the U.S. industrial sector as measured by the real share of IT equipment and software in the real aggregate capital stock.[3]

[3] We build the ratio plotted in Figure 3 by summing the capital stocks of 62 industrial sectors from the detailed non-residential fixed asset tables in fixed 1996 dollars made available by the Bureau of Economic Analysis (2002). IT capital includes mainframe and personal computers, storage devices, printers, terminals, integrated systems, and pre-packaged, custom, and own-account software. The total capital stock is the sum of all fixed asset types.

Computer and software purchases appear to have reached the first inflection point in their "S-curve" more slowly than Electrification in the early years of its GPT-era, but it is striking how much faster the IT share has risen over the past few years. Moreover, while the diffusion of electricity had slowed down by 1930, the year which we mark as the end of the Electrification "era," yet computer and software sales continue their rapid rise to this day.

The scaling of the vertical axes in Figure 2 and Figures 3 is different. The vertical axis in Figure 2 measures the shares of total horsepower in manufacturing, whereas the vertical axis in Figures 3 is the real share of IT equipment and software in the real aggregate capital stock. But scaling aside, the diffusion process appears to be much more protracted for IT than for Electrification insofar as a comparison of the shape of the diffusions in the two figures suggests that the IT era will last longer than the 35 years of Electrification. The acceleration in adoption, which was over by about 1905 for Electrification, did not end until about 1997 for IT.

Why did electricity spread faster than IT seems to be doing? We do not know if this is because it was more profitable, or because the rapid price decline of computers and peripherals makes it optimal to wait and adopt later, or still for some other reson.

2.1.2 Pervasiveness among sectors

Cummins and Violante (2002, p. 245) classify a technology as a GPT when the share of new capital associated with it reaches a critical level, and that adoption is widespread across industries. Electrification seems to fit this description. Figure 4 shows the shares of total horsepower electrified in manufacturing sectors at ten-year intervals from 1889 to 1954.[4] Electrical adoption was very rapid between 1899 and 1919 but slowed considerably thereafter, with the dispersion in the adoption rates largest around 1919.

The main message in Figure 4 is that electrical technology affected individual manufacturing sectors with a striking degree of uniformity. Moreover, Table 1, which shows the rank correlations of electricity shares across sectors and time, indicates that there was little change in the relative ordering of the manufacturing sectors either. This means that the sectors that were the heaviest users of electricity in 1890 remained among the leaders as adoption slowed down in the 1930s. Indeed, the adoption of electricity was sweeping and widespread.

It was not practical to set up the wiring required to electrify households early on. This is apparent from the peculiar two-stage adoption process that many factories chose in adopting electricity: Located to a large extent in New England factory towns, textile firms around the turn of the century readily adapted the new technology by using an electric motor rather than steam to drive the shafts which powered looms,

[4]The shares of electrified horsepower include primary and secondary electric motors, and are computed using data from DuBoff (1964, tables E-11 and E-12a through E-12e, pp. 228-235).

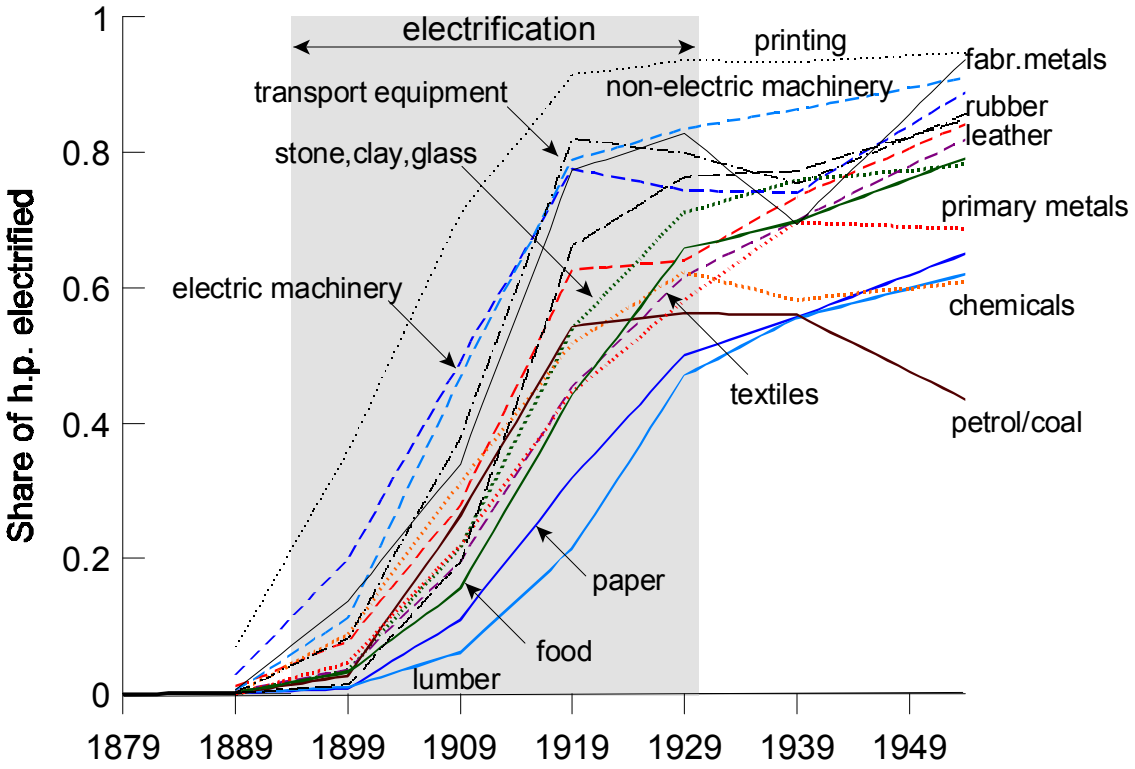


Figure 4: Shares of electrified horsepower by manufacturing sector, 1890-1954.

spinning machines and other equipment [see Devine (1983)]. Further delays in the distribution of electricity made it more costly to electrify a new industrial plant fully.

Table 1
Rank correlations of electricity shares in total horsepower
by manufacturing sector, 1889-1954

	1889	1899	1909	1919	1929	1939	1954
1889	1.000						
1899	0.707	1.000					
1909	0.643	0.918	1.000				
1919	0.686	0.746	0.893	1.000			
1929	0.639	0.718	0.739	0.871	1.000		
1939	0.486	0.507	0.571	0.750	0.807	1.000	
1954	0.804	0.696	0.650	0.789	0.893	0.729	1.000

Figure 5 shows the same data as Figure 4, but now in percentile form. We build them by sorting the electricity shares in each year and, given that only 15 sectors

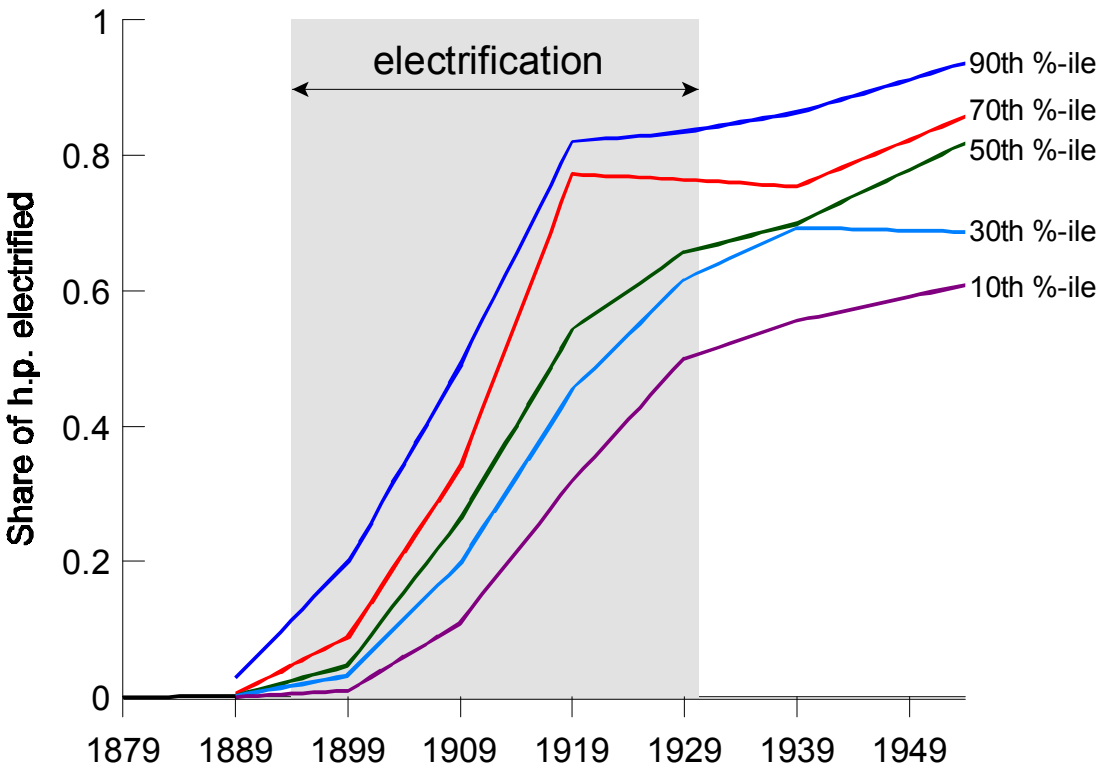


Figure 5: Shares of electrified horsepower by manufacturing sector in percentiles, 1890-1954.

are represented, plotting the 2nd, 5th, 8th, 11th and 14th largest shares in each year. The percentile diffusion curves will be useful when drawing comparisons with the IT era. They also help us in dating electricity as a GPT. Linear extrapolation between the years 1890 and 1900 suggests that in 1894, about one percent of horsepower equivalents in the median industry was provided by electricity. Whether or not this is the right percentage for dating the start of the Electrification era, we shall use the same percentage for the median industry to date the beginning of the IT era, thereby using a common standard for choosing the left-end points of the two shaded areas.

In the century before the Electricity revolution, the technology that primarily drove manufacturing was steam. Figure 6 shows just how slowly steam was replaced between 1899 and 1939.[5] It is natural that industries such as rubber, primary metals, non-electric machinery, and stone, clay, and glass, which saw such rapid increases in electricity use over the same period, would withdraw from steam most rapidly.

[5]The sectoral shares of manufacturing horsepower driven by steam were computed from DuBoff (1964, tables E-12a through E-12e, pp. 229-233), and include steam engines and turbines. These shares are available on a decade basis from 1899 until 1939 only, which is why the time coverage in Figure 6 is shorter than that in Figure 4.

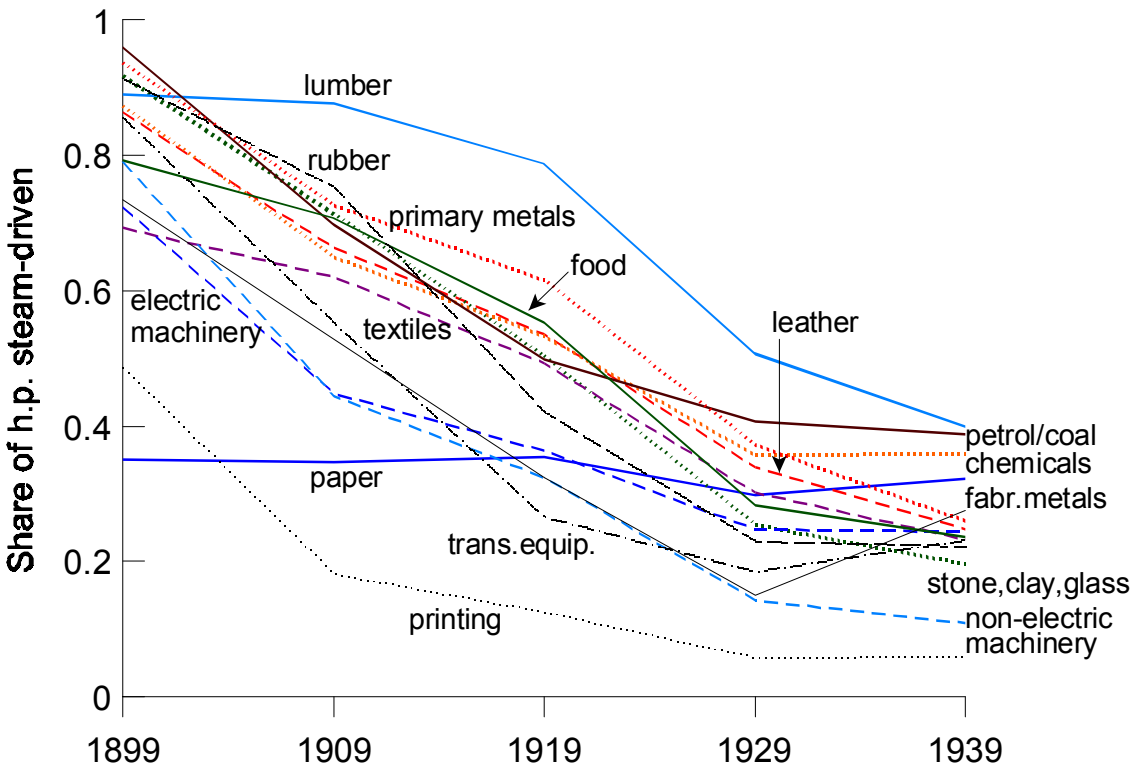


Figure 6: Shares of steam-driven horsepower by manufacturing sector, 1899-1939.

Indeed, most of the industries that quickly switched over to electricity had been heavy users of steam. This is clear from Figure 4 and Figure 6, taken together, and from the rank correlations presented in Table 2 which decay quickly and suggest a non-uniformity in the destruction of steam technology across sectors.

Table 2
Rank correlations of steam shares in total horsepower
by manufacturing sector, 1889-1954

	1899	1909	1919	1929	1939
1899	1.000				
1909	0.825	1.000			
1919	0.604	0.800	1.000		
1929	0.525	0.604	0.832	1.000	
1939	0.261	0.282	0.496	0.775	1.000

The spread of information technology was also rapid, but does not appear to have been as widespread as electricity. Figure 7 shows the share of real IT equipment and software in the real net capital stocks of 62 sectors from 1960 to 2001 plotted as

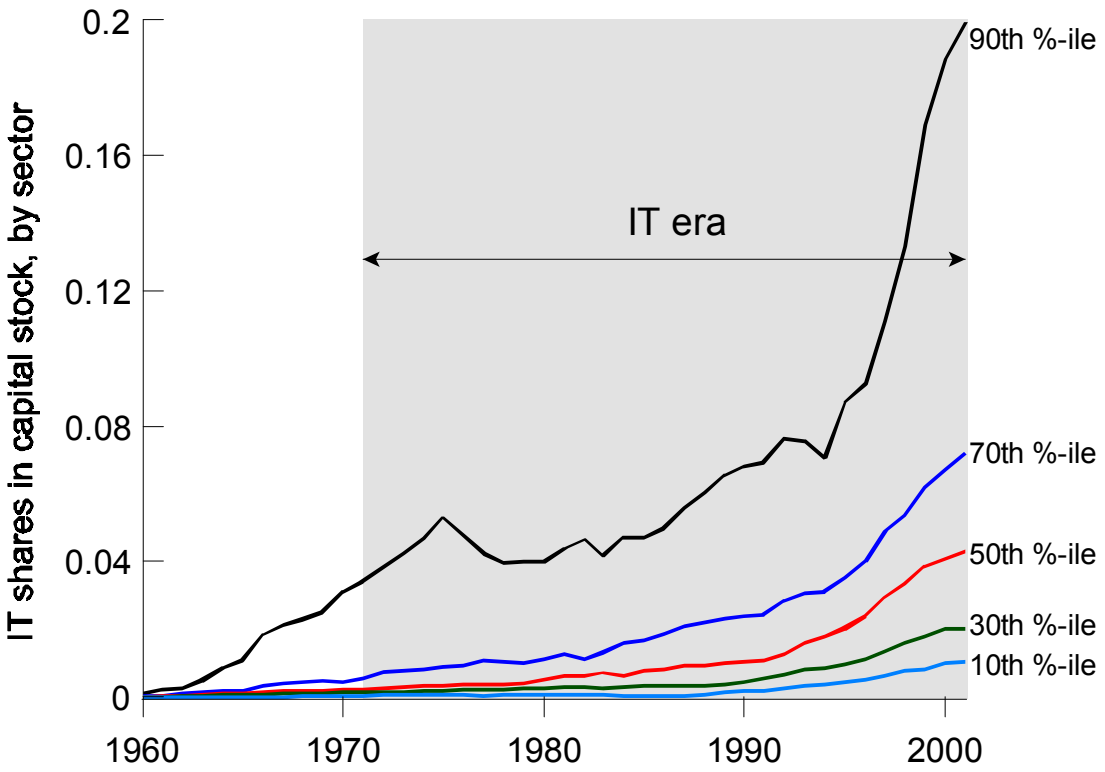


Figure 7: Shares of IT equipment and software in the capital stock by sector in percentiles, 1960-2001.

annual percentiles.[6] In the case of IT, some sectors adopted very rapidly, and by 1975 six of them (the 90th percentile) had already achieved IT equipment and software shares of more than 5 percent. Other sectors lagged behind, and some did not adopt IT in a substantive way until after 1985.

On the other hand, the rank correlations of the IT shares across sectors, shown in Table 3, are even higher than those obtained for Electrification. On the face of it then, Electrification would appear to have been the more sweeping GPT-type event because it diffused more rapidly in the U.S. economy and all sectors adopted it pretty much at the same time, whereas IT diffused rapidly in some sectors and not-so-rapidly in others. Nonetheless, the recent gains in IT shares show that the diffusion of this GPT has yet to slow down in the way that Electrification did after 1929.

[6]The sectoral capital stocks are from the detailed non-residential fixed asset tables in fixed 1996 dollars made available by the Bureau of Economic Analysis (2002). We present the sectoral shares for the IT epoch in percentile form because the number of sectors covered is much larger than was possible for electrification and steam.

Table 3
Rank correlations of IT shares in capital stocks
by sector, 1961-2001

	1961	1971	1981	1991	2001
1961	1.000				
1971	0.650	1.000			
1981	0.531	0.806	1.000		
1991	0.576	0.746	0.847	1.000	
2001	0.559	0.682	0.734	0.909	1.000

So far we have discussed adoption by firms, and it determines the dating of the two GPT epochs. We turn to households next.

2.1.3 Adoption by households

Households also underwent electrification and the purchase of personal computers for home use during the respective GPT-eras. Figure 8 shows the cumulative percentage of households that obtained electric service and that owned a personal computer in each year following the "arrival" of the GPT.[7] If we continue to date Electricity as arriving in 1894 and the personal computer in 1971, Figure 8 shows that households adopted electricity about as rapidly as they are adopting the personal computer. By the time the technology is officially 35 years old (in 1929), nearly seventy percent of households had electricity. A comparison with Figure 5 shows that this is just a little higher than the 1929 penetration of electrified horsepower share by the median manufacturing sector. As in the case of firms, the Electrification of households reaches a plateau in 1929, although it resumed its rise a few years later. On the other hand, there is no sign yet that the diffusion of the computer among either households or firms is slowing down.

With households, as with firms, diffusion lags seem to arise for different reasons for the two technologies. Rural areas were difficult to reach for Electricity, but not so for the PC, for which the main barrier is probably the cost of learning how to use it. This barrier seems to have more to do with human capital than was the case with Electricity.

In some ways it is puzzling that the diffusion of the PC has not been much faster than that of Electrification. The price of computing is falling much faster than the price of Electricity did. Affordable PCs came out in the 1980s, when the technology was some 15 years old. On the other hand, households had to wait longer for affordable electrical appliances. Only after 1915, when secondary motors begin to spread widely,

[7]Data on the spread of electricity use by consumers are approximations derived from U.S. Bureau of the Census (1975) *Historical Statistics of the United States* (series S108 and S120). Statistics on computer ownership for 1975 through 1998 are from Gates (1999, p. 118), and from the Census Bureau's *Current Population Survey* thereafter.

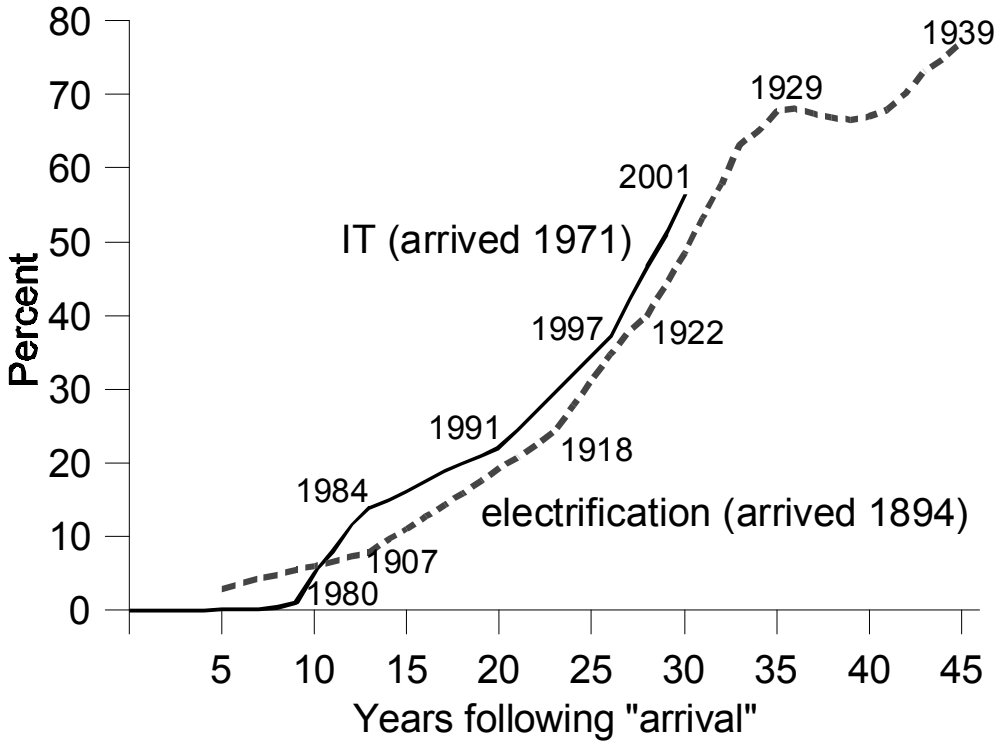


Figure 8: Percent of households with electric service and personal computers during the two technological epochs.

and electrical appliances began to be invented, did the benefits of Electrification outweigh the costs for a majority of households. Greenwood, Shesadri and Yorukoglu (2001) document the spread of electrically powered household appliances and argue that their diffusion helped raise female labor-force participation by freeing up their time from housework.

2.1.4 On dating the endpoints of the era of a GPT

Our dating procedure reflects net adoption rates by firms, but the dates would not change much if we had instead used net adoption by households. The shaded areas are periods when the S-shaped adoption curves are, for the most part, rising. Whether or not they start to fall later should not affect the designated adoption eras. For instance, electricity has not yet been replaced in the same way that steam was phased out in the first half of the twentieth century, but the "Electrification era" still ends in 1930 because adoption as measured in Figures 2, 4, and 5 flattens out. Figures 2 and 6 show that the steam era must have ended sometime around 1899 because net adoption is already negative.

Net adoption is endogenous and it should reflect the profitability of the technology at hand compared to that of other technologies. The Niagara Falls dam in 1894 and the development of alternating current made it possible to produce and distribute electricity more cheaply at greater distances. Figures 4 and 5 show that at the outset, some sectors (like printing) raced ahead of others in terms of how quickly they adopted. Later on, as the technology matures, its adoption becomes more universal. Eventually, the lagging sectors tend to catch up a bit, in relative terms, but not completely. Inequality of adoption is highest in the middle of the adoption era. We also see such a temporary rise in inequality in Figure 6 about the same time.

2.2 Improvement of the GPT

The second characteristic that Bresnahan and Trajtenberg suggested is an improvement in the efficiency of the GPT as it ages. Presumably this would show up in a decline in prices, an increase in quality, or both. How much a GPT improves can be therefore measured by how much cheaper a unit of quality gets over time. If technology is embodied in capital, then presumably, capital as a whole should be getting cheaper faster during a GPT era, but especially capital that is tied to the new technology.

To answer these questions, we first look at the prices of capital as a whole and then at the prices of its components. Figure 9 is a quality-adjusted series for the relative price of equipment as a whole, p_k/p_c (i.e., relative to the consumption price index), since 1885, constructed from a number of sources, with a linear time trend included.[8] The figure shows that equipment prices declined most sharply between 1905 and 1920, and again after 1975. The 1905-1920 period is also the one that showed the most rapid growth of electricity in manufacturing (see Figure 4) and in the home (see Figure 8). The post-1975 period follows the introduction of the PC.

Figure 10 aims to look at the components of the aggregate capital stock; specifically, the components tied to the two GPTs. Because deflators for electrically powered capital are not available in the first half of the twentieth century, Figure 10 compares the declines in relative prices associated with three GPT's – electricity, internal com-

[8] Krusell et al. (2000) build such a series from 1963 using the consumer price index to deflate the quality-adjusted estimates of producer equipment prices from Gordon (1990, table 12.4, col. 2, p. 541). Since Gordon's series ends in 1983, they use VAR forecasts to extend it through 1992. We start with Krusell et al. and work backward, deflating Gordon's remaining estimates (1947-62) with an index for non-durable consumption goods prices that we derive from the National Income Accounts. Since we are not aware of a quality-adjusted series for equipment prices prior to 1947, we use the average price of electricity as a proxy for 1902-46, and an average of Brady's (1966) deflators for the main classes of equipment for 1885-1902. We deflate the pre-1947 composite using the Bureau of Labor Statistics (BLS) consumer prices index of all items [U.S. Bureau of the Census (1975, series E135)] for 1913-46 and the Burgess cost of living index [U.S. Bureau of the Census (1975, series E184)], which has greater precision than the BLS series, for 1885-1912.

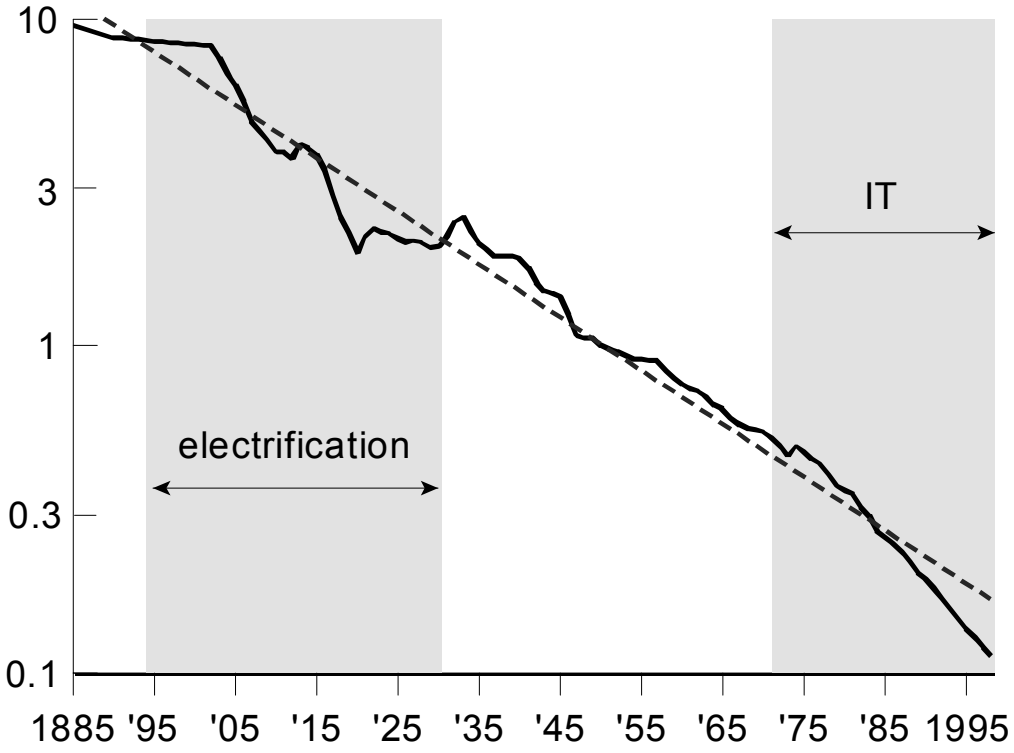


Figure 9: The relative price of equipment.

bustion, and computers – once again relative to the consumption price index.[9] The

[9]To construct a quality-adjusted price index, we join the "final" price index for computer systems from Gordon (1990, table 6.10, col. 5, p. 226) for 1960-78 with the pooled index developed for desktop and mobile personal computers by Berndt, Dulberger, and Rappaport (2000, table 2, col. 1, p. 22) for 1979-99. Since Gordon's index includes mainframe computers, minicomputers, and PCs while the Berndt et al. index includes only PCs, the two segments used to build our price measure are themselves not directly comparable, but a joining of them should still reflect quality-adjusted prices trends in the computer industry reasonably well. We set the index to 1000 in the first year of the sample (i.e., 1960).

Electricity prices are averages of all electric energy services in cents per kilowatt hour from the *Historical Statistics of the United States* (U.S. Bureau of the Census, 1975, series S119, p. 827) for 1903, 1907, 1917, 1922, and 1926-70, and from the *Statistical Abstract of the United States* for 1971-89. We interpolate under a constant growth assumption between the missing years in the early part of the sample. For 1990-2000, prices are U.S. city averages (June figures) from the Bureau of Labor Statistics (*http:www.bls.gov*). We then set the index to equal 1000 in the first year of the sample (i.e., 1903).

Motor vehicle prices for 1913-40 are annual averages of monthly wholesale prices of passenger vehicles from the National Bureau of Economic Research (Macrohistory Database, series m04180a for 1913-27, series m04180b for 1928-40, *http://www.nber.org*). From 1941-47, they are wholesale prices of motor vehicles and equipment from *Historical Statistics* (series E38, p. 199), and from 1948-2000 they are producer prices of motor vehicles from the Bureau of Labor Statistics (*http://www.bls.gov*).

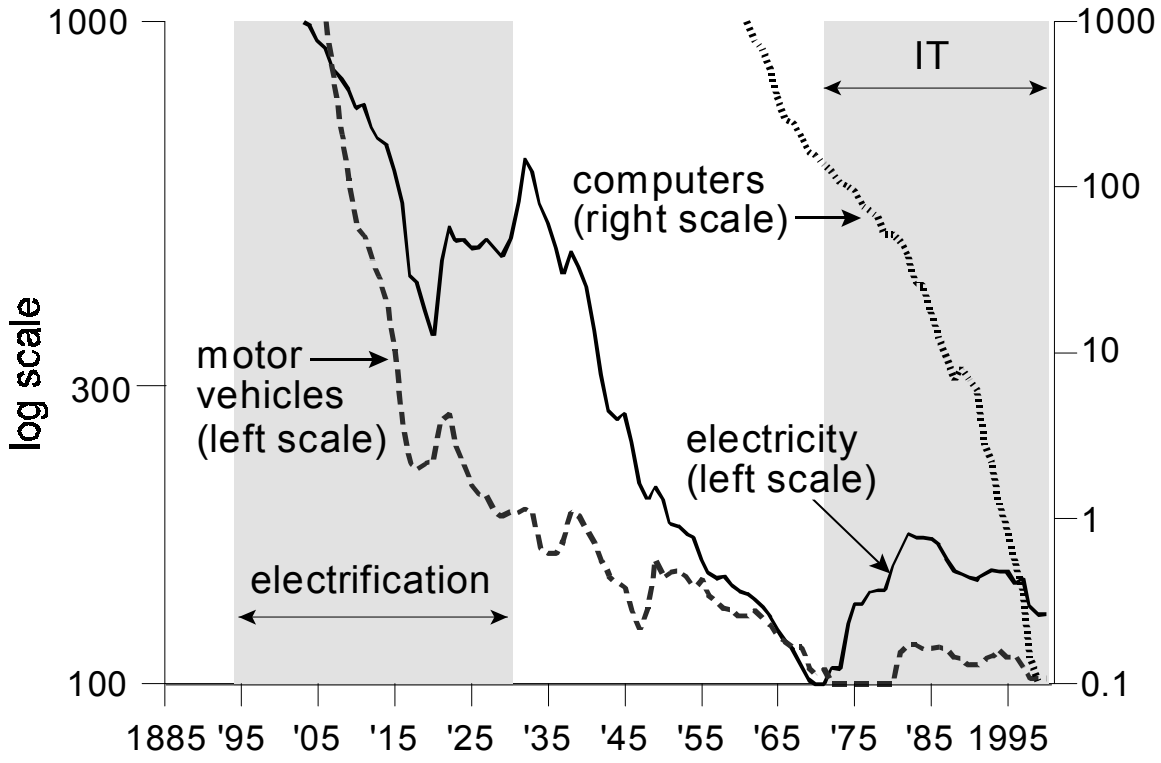


Figure 10: Price indexes for products of two technological revolutions.

use of the left-hand scale for electricity and motor vehicles and the right-hand scale for computers underscores the extraordinary decline in computer prices since 1960 compared to the earlier technologies. While the electricity and the automobile indexes fall by a factor of 10, the computer index falls by a factor of 10,000.

The more important question, however, is how the general decline in equipment prices compares to the declines associated more directly with the GPTs of each epoch. Figure 11 makes this comparison by plotting the relative prices of all three GPT's along with the general equipment index on the same logarithmic scale, with the starting point for each of the GPTs normalized to the level of the general equipment price index in that year. By this measure, it is clear that electricity and motor vehicle prices declined at about the same pace as that of equipment generally until the start of the IT price data, though it is also interesting that motor vehicle prices appear to have declined faster than electricity prices. After 1960, declining computer prices and rising shares of computers in equipment stocks seem to have drawn the general index downward, while computing prices fell thousands of times faster than the general

To approximate prices from 1901-1913, we extrapolate assuming constant growth and the average annual growth rate observed from 1913-24. We then join the various components to form an overall price index, and set it to equal 1000 in the first year of the sample (i.e., 1901).

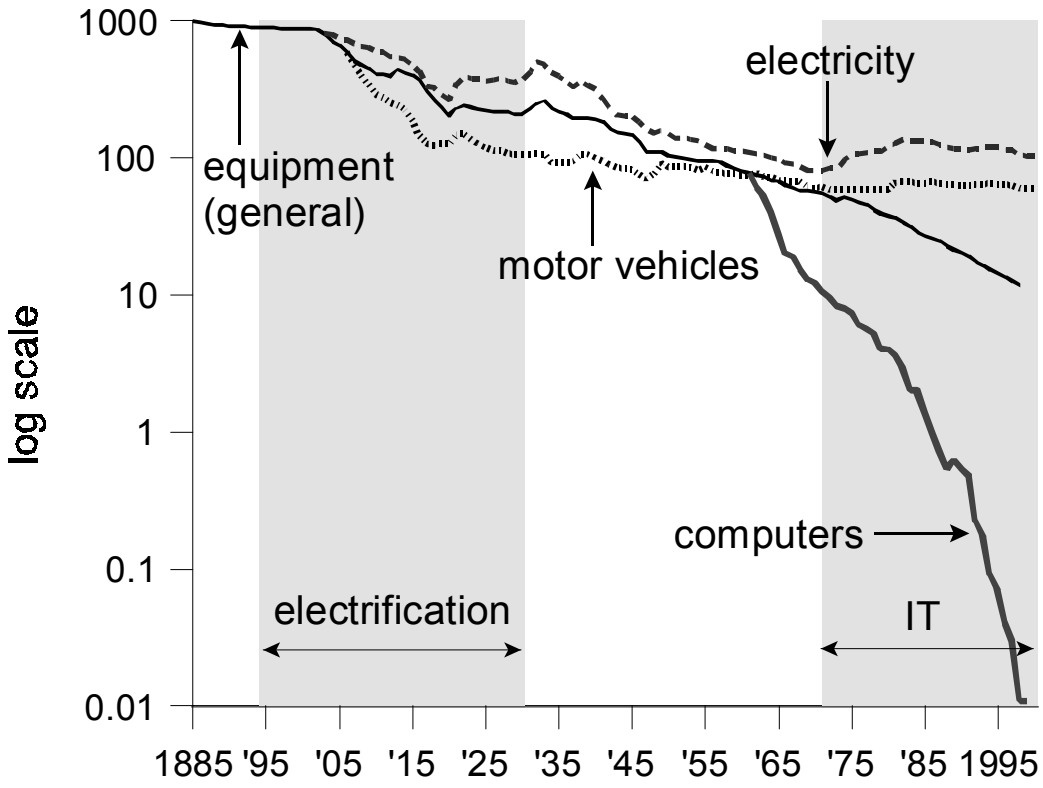


Figure 11: Comparison of the decline in general and GPT-specific equipment prices.

equipment index.

It can be said that the electricity index, being the price of a kilowatt hour, understates the accompanying technological change because it does not account for improvements in electrical equipment, and especially improvements in the efficiency of electrical motors. Such improvements may be contained in the price series for capital generally. Yet based on the price evidence in figures 10 and 11, electricity, motor vehicles, and computers might all qualify as GPTs, but computers are clearly the most revolutionary of the three.

2.3 Ability of the GPT to spawn innovation

The third characteristic that Bresnahan and Trajtenberg suggested was the technology's ability to generate innovation. Any GPT will affect all sorts of production processes, including those for invention and innovation. Some GPTs will be biased towards helping to produce existing products, others towards inventing and implementing new ones. An example of a more specific technology that was heavily skewed towards future products was hybrid corn. Griliches (1957, p. 502) explains why hy-

brid corn was not an invention immediately adaptable everywhere, but, rather, that it was the invention of a method of inventing, a method of breeding superior corn for specific localities.

Electricity and IT have both helped reduce costs of making existing products, and they both spawn innovation, but IT is more skewed towards the latter. The 1920s especially saw a wave a new products powered by electricity, and the computer is now embodied in many new products as well. But as the patenting evidence will bear out, IT seems to have more of a skew towards contributing to further innovation – the role of the computer in simulation should be known to many of us writing research papers. Feder (1988) describes how computers play a similar role in the invention of new drugs.

2.3.1 Patenting

Patenting should be more intense after a GPT arrives and while it is spreading due to the introduction of related new products. Figure 12, which shows the per capita numbers of patents issued on inventions annually from 1790 to 2000 and trademarks registered from 1870 to 2000, shows two surges in activity – between 1900 and 1930, and again after 1977.[10] Is it mere chance that patenting activity was most intense during our technological revolutions? Moreover, it appears that patenting activity picks up after the end of the U.S. Civil War in 1865, and again at the conclusion of World War II in 1945. The slowdown in patenting during the wars and acceleration immediately thereafter suggest that there is some degree of intertemporal substitution in the release of new ideas away from times when it might be more difficult to popularize them and towards times better suited for the entry of new products.

Does the surge in patenting reflect a rise in the number of actual inventions, or was the surge prompted by changes in the law that raised the propensity to patent? This question is important because, over longer periods of time, patents may reflect policy rather than invention: Figure 13 analyzes data described in detail in Lerner (2002) and shows that worldwide, changes in patent policy changes are correlated with the patent series in Figure 12. It is possible, therefore, that the U.S. series reflects court-enforcement attitudes. Kortum and Lerner (1998) analyze this question and found that the surge of the 1990s was worldwide, but not systematically related to country-specific policy changes, and they conclude that technology was the cause for the surge.

[10] We use the total number "utility" (i.e., invention) patents from the U.S. Patent and Trademark Office for 1963-2000, and from the U.S. Bureau of the Census (1975, series W-96, pp. 957-959) for 1790-1962. The number of registered trademarks are from the U.S. Bureau of the Census (1975, series W-107, p. 959) for 1870-1969, and from various issues of the *Statistical Abstract of the United States* for later years. Population figures, which are for the total resident population and measured at mid-year, are from U.S. Bureau of the Census (1975, series A-7, p. 8) for 1790-1970, and from the Bureau of Economic Analysis thereafter.

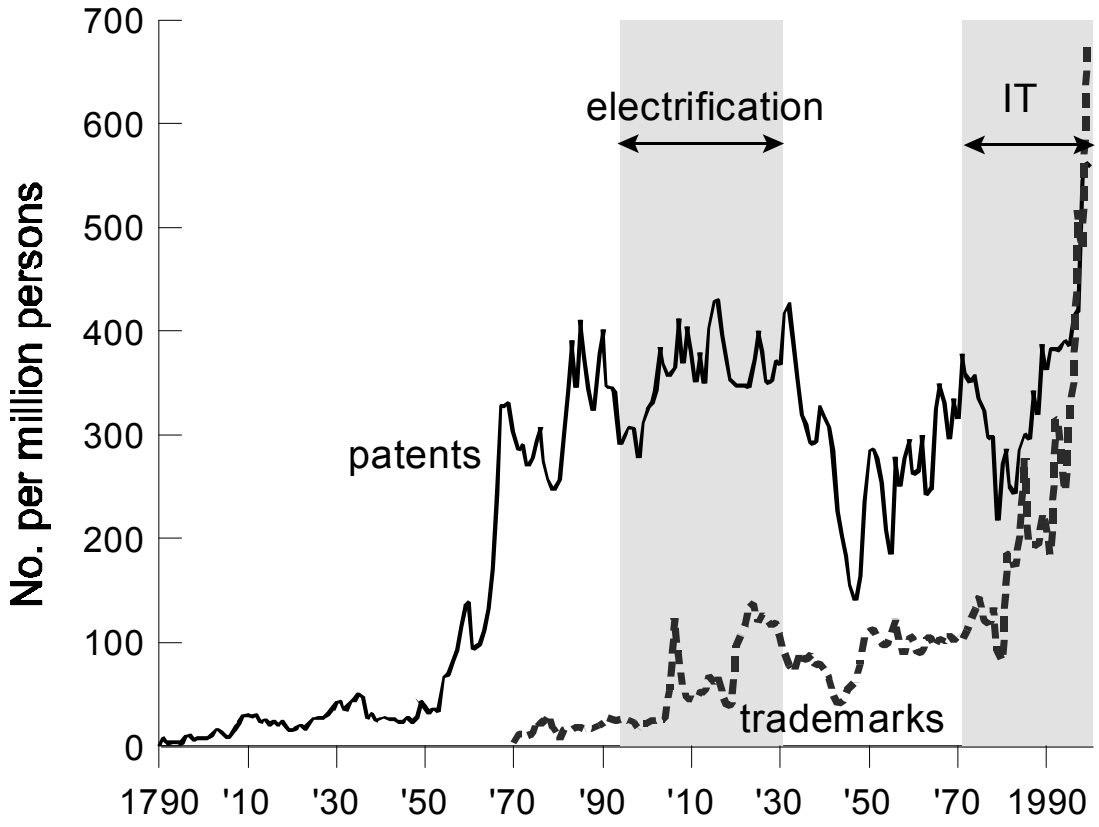


Figure 12: Patents issued on inventions and trademarks registered in the United States per million persons, 1790-2000.

Further support for this view comes from the behavior of trademarks per head, which we also plot in figure 12. Trademarks behave more or less the same as patents do, except for their higher trend. Trademarks are easier to obtain than patents and are not governed by legal developments concerning patents. But with trademarks we have a different concern: Do trademarks proxy for the number of products, or do they just measure duplicative activity and the amount of competition? The answer may depend on what market one looks at. In the market for bananas, for example, Wiggins and Raboy (1996) find that brand names are correlated with measures of quality that do explain price variation, suggesting, therefore, that brand names do signify product differentiation.

2.3.2 Investment in new firms vs. investment by incumbents

If new technologies are more easily embraced by new firms that do not bear the burden of costs sunk in old technologies and the rigid and firm-specific organization capital required to operate them, we should expect to see waves of new listings on

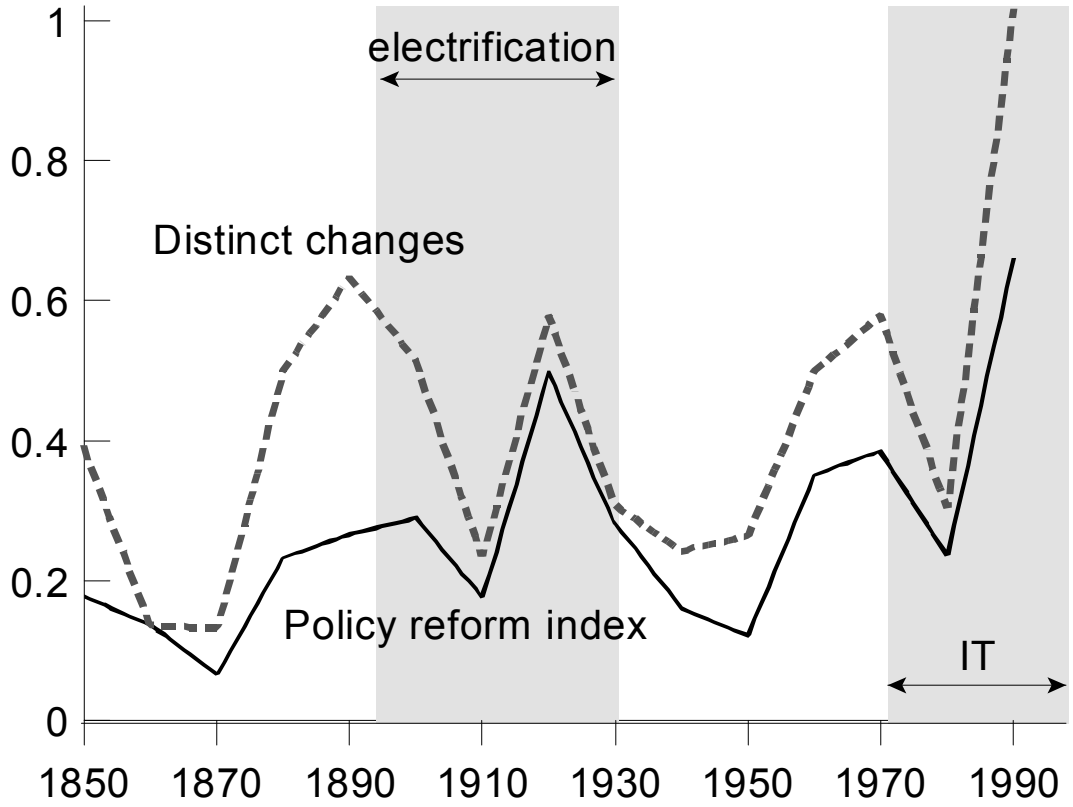


Figure 13: Indexes of worldwide changes in patent laws.

the stock exchange during GPT eras. Figure 14 shows the value of firms entering the New York Stock Exchange (NYSE), the American Stock Exchange (AMEX), and NASDAQ in each year from 1885 through 2001 as percentages of total stock market value.[11] As predicted by Trajtenberg and Helpman, IPOs surge between 1895 and 1929, and then after 1977, which again closely matches the dating of our two technological revolutions.

[11]The data used to construct Fig. 14 and others in this chapter that use stock market valuations are from the University of Chicago's Center for Research in Securities Prices (CRSP) files for 1925-2001. NYSE firms are available in CRSP continuously, AMEX firms after 1961, and NASDAQ firms after 1971. We extended the CRSP stock files backward from their 1925 starting year by collecting year-end observations from 1885 to 1925 for all common stocks traded on the NYSE. Prices and par values are from the *The Commercial and Financial Chronicle,* which is also the source of firm-level data for the price indexes reported in the well-known study by Cowles et al. (1939). We obtained firm book capitalizations from *Bradstreet's*, *The New York Times*, and *The Annalist*. The resulting dataset, which includes 24,475 firms, though limited to annual observations, actually includes more common stocks than the CRSP files in 1925. As such, the dataset complements others that have begun to build a more complete view of securities prices in other markets for the pre-CRSP period [see, for example, Rousseau (1999) on Boston's 19th century equity market.

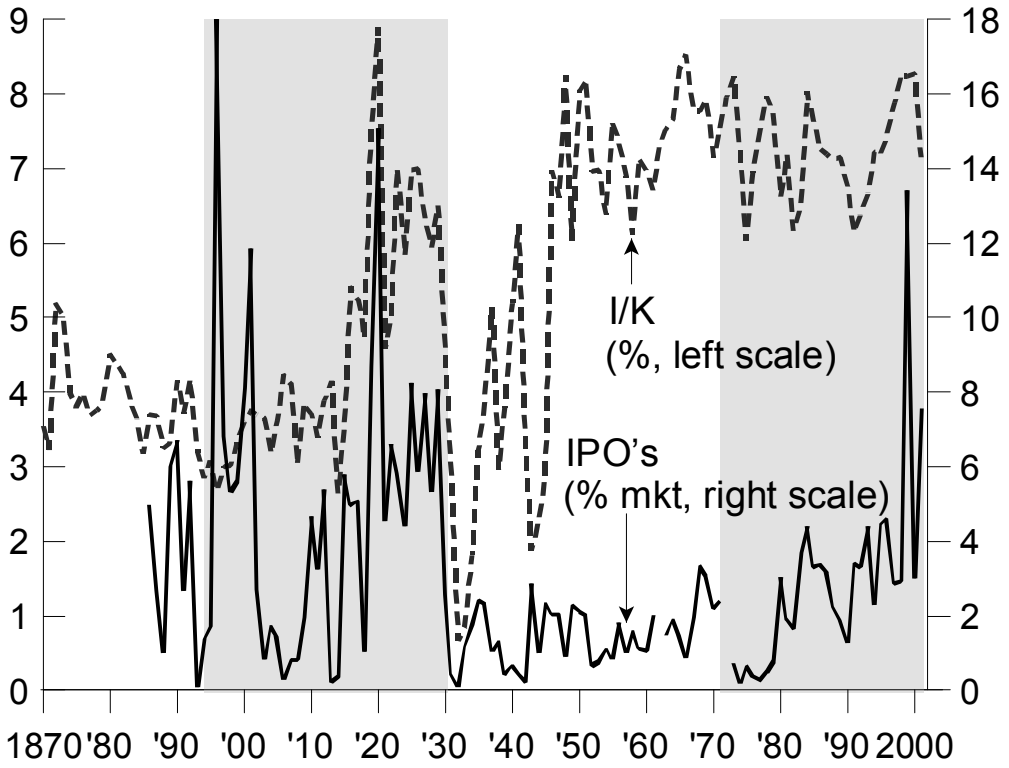


Figure 14: Annual IPOs as a percent of stock market value, and investment as a percent of the net capital stock, 1870-2001.

The dashed line in Figure 14 is private investment since 1870 as a percent of the net stock of private capital for the U.S. economy as a whole, and as such is the aggregate analog of the solid line that covers only the stock market.[12] The solid line in Figure 15 shows the ratio of the solid and dashed lines in Figure 14. In both figures it is clear that, during Electrification, investment by stock market entrants accounted for a larger portion of stock market value than overall new investment in

[12]To build the investment rate series, we start with gross private domestic investment in current dollars from the Bureau of Economic Analysis (2002, Table 1, pp. 123-4) for 1929-2001 and then ratio splice the gross capital formation series in current dollars, excluding military expenditures, from Kuznets (1961b, Tables T-8 and T-8a) for 1870-1929. We construct the net capital stock using the private fixed assets tables of the Bureau of Economic Analysis (2002) for 1925-2001. Then, using the estimates of the net stock of non-military capital from Kuznets (1961a, Table 3, pp. 64-5) in 1869, 1879, 1889, 1909, 1919, and 1929 as benchmarks, we use the percent changes in a synthetic series for the capital stock formed by starting with the 1869 Kuznets (1961a) estimate of $27 billion and adding net capital formation in each year through 1929 from Kuznets (1961b) to create an annual series that runs through the benchmark points. Finally, we ratio-splice the resulting series for 1870-1925 to the later BEA series. The investment rate that appears in Figure 14 is the ratio of our final investment to the capital stock series, expressed as a percentage.

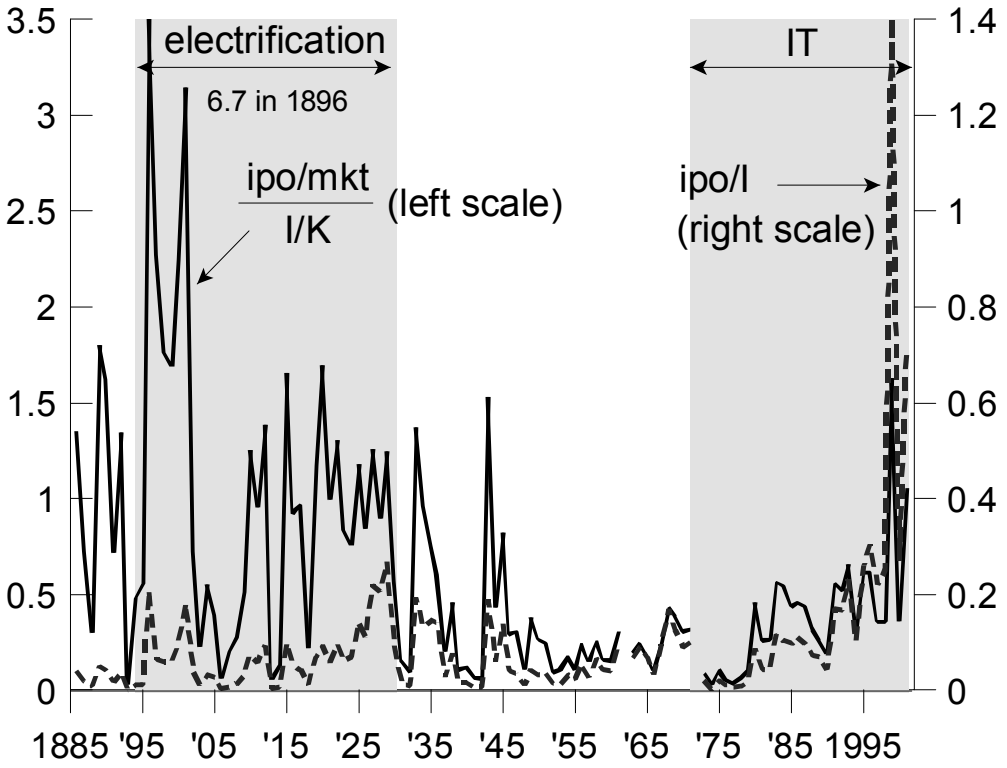


Figure 15: Ratio of IPOs as a percent of the stock market to investment as a percent of the net capital stock, 1885-2001.

the U.S. economy contributed to the aggregate capital stock. This is consistent with the adoption of electricity favoring the unencumbered entrant over the incumbent, who may have incurred large adjustment costs in using the new technology. We say this because aggregate investment, while indeed including new firms, has an even larger component attributable to incumbents. Moreover, the ratio of IPO to aggregate investment activity was highest in the early years of Electrification, which is when these adjustment costs would have been greatest. Although the ratio given by the solid line in Figure 15 has so far stayed below unity for most of the IT-era, it has been rising rapidly in recent years. This could be because IT adoption involved very large adjustment costs for both incumbents and entrants in the early years until the price of equipment and software fell enough to promote adoption among new firms.

The solid line in Figure 15 shows a downward trend mainly because the stock market gained importance as a model of financing in the early part of the 20th century. IPOs are normalized by the stock market which was small early on, and has since become larger. The dotted line in Figure 15 shows the ratio of the un-normalized series of IPOs and aggregate investment. This has a positive trend for

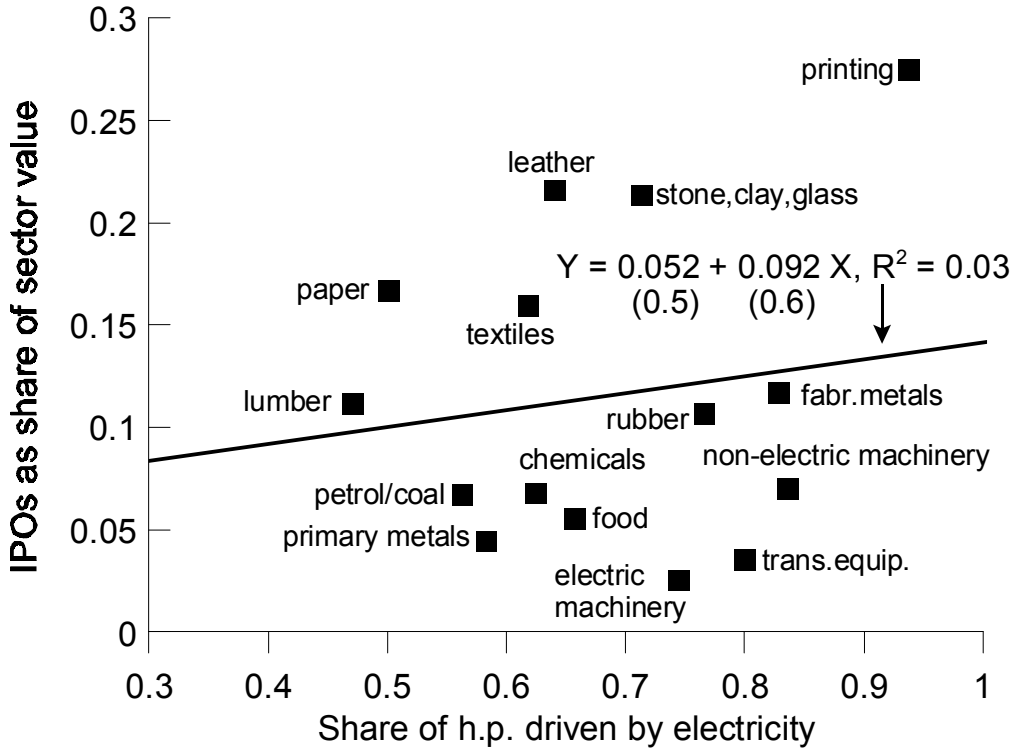


Figure 16: IPOs as shares of sector value 1890-1930 vs. shares of electricity-driven horsepower in 1929.

the same reason that the previous figure showed a negative one: IPOs were not that important early on because the stock market was small. After 1970, IPOs capture a much larger share of the investment by new entrants than they did before the first World War, for example, and even a larger fraction than in the 1920s. When we consider both lines together, we do get the impression that new firms invest more during the GPT eras than at other times.

Does the distribution of entries across sectors shed light on the role of technological factors in the entry waves? Perhaps so. Figure 16 is a scatterplot of the share of IPOs in the market capitalization of 15 manufacturing sectors between 1890 and 1930 vs. their respective shares of horsepower driven by electricity in 1929.[13] In other words, we ask whether sectors with more IPOs ended up embracing the new technology more vigorously than sectors with less entry. The regression line plotted in Fig. 16

[13]We compute the IPO shares by sector by summing year-end IPO values by sector for 1890 through 1930, converting each year's total into real terms using the implicit price deflator for gross domestic product, and then summing across years. We do the same for all listed firms by sector, and use the quotient of sector IPO values and total sector values to compute the shares shown in Fig. 16.

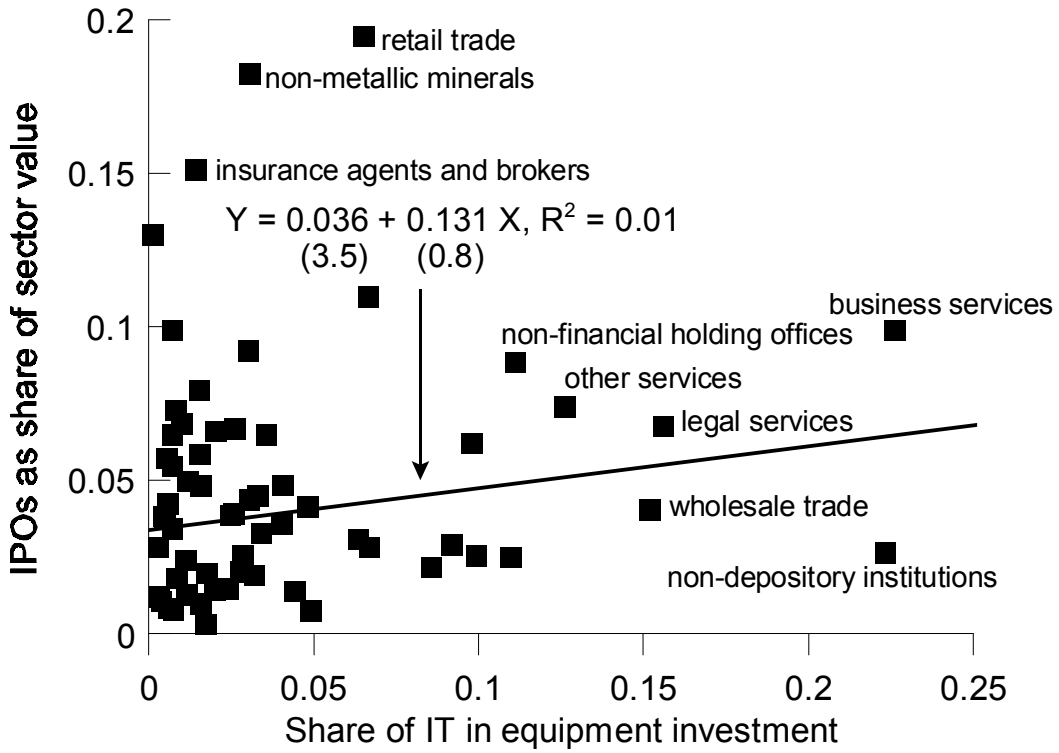


Figure 17: IPOs as shares of sector value 1971-2001 vs. shares of IT in equipment investment in 2000.

indicates that this relationship is indeed positive, though with only 15 observations the slope coefficient is not statistically significant at conventional levels.

In Figure 17, we consider IPOs over the 1971-2001 IT-epoch against shares of computers and peripherals in equipment investment in 2000, we once again obtain a sectoral scatter with a positive slope, though like Electrification, the slope coefficient is not statistically significant.

3 Other symptoms of GPT

So far we have provided some measures of the three qualities of a GPT – its pervasiveness, its rate of improvement, and its innovation-spawning tendency. Now we turn to less direct measures, as are suggested by various theoretical models that deal with GPTs and that predict the following symptoms:

1. *Productivity should slow down.*—The new technology may not be user friendly at first, and output may fall for a while as the economy adjusts.

2. *The skill premium should rise.*—If the GPT is at first user unfriendly, skilled people will be in greater demand when a new technology arrives, and their earnings should rise compared to those of the unskilled.

3. *Entry, exit and mergers should rise.*—These are alternative modes of reallocation of assets.

4. *Stock prices should initially fall.*—The value of old capital should fall; how fast it does so depends on the way that the market learns of the GPT's arrival.

5. *Young and small firms should do better.*—The ideas and products associated with the GPT will often be brought in by new firms. The market share and market value of young firms should rise relative to old firms

6. *Interest rates and the trade deficit.*—The rise in desired consumption relative to output should cause interest rates to rise or the trade balance to worsen.

These are not propositions but hypotheses that one can find in much of the work on GPTs. Now we give evidence on each one in turn. Each gets its own subsection.

3.1 Productivity slowdown

Even in routine activities, learning seems to cause delays of several years before plant productivity peaks [Bahk and Gort (1993)]. It is far from settled whether IT is the reason for the productivity slowdown – Bessen (2002) finds that IT did cause a big part of the slowdown, whereas Comin (2002) argues the opposite. It also is not yet definitely known from the work of Caballero and Hammour (1994) and others whether recessions at business-cycle frequencies are episodes of heightened reallocation. At any rate, the theoretical models of Atkeson and Kehoe (1993), Hornstein and Krusell (1996), Jovanovic and Nyarko (1996), Greenwood and Yorukoglu (1997) and Jovanovic and Rousseau (2002a) emphasize various adjustment costs and learning delays that may cause output to fall at first when a GPT arrives. David (1991) argues that the speed with which a new technology diffuses depends on the pool of investment opportunities that are available when it arrives, and remarks that the quality of this pool in the late 1960s was low because the large backlog of the post-war period had just and finally been eliminated. He also points out that there can often be "slippage" between the technology frontier and implementation due to high input costs and the slow introduction of complementary products.

Figure 1 shows that productivity does not rise quickly in the early phases of the two GPTs, though there is some evidence of greater productivity between 1918 and 1929. This could be taken as further evidence that IT did not do as much as did electricity to raise the productivity on existing processes and products. Productivity was high in the early years of Electrification but fell rapidly as the technology matured. It

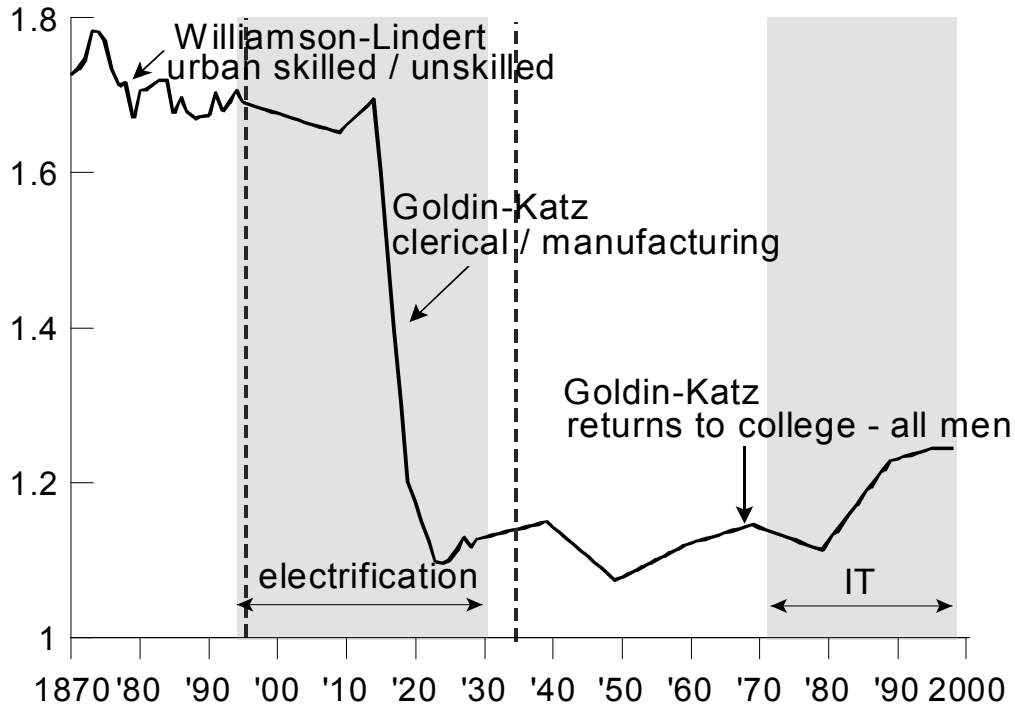


Figure 18: The skill premium.

stayed low through the Depression and 1940s, and then rose rapidly before the IT-age arrived. This pattern is consistent with David's model of exhausted investment opportunities. And while it is interesting to consider the productivity slowdown after 1971, it is also important to recognize that productivity is considerably higher today than it was before IT's arrival.

3.2 The skill premium

As Nelson and Phelps (1966) and Griliches (1969) argued, and Bartel and Lichtenberg (1987) and Krusell et al. (2000) have confirmed, new technology should raise the relative earnings of the skilled. Figure 18 presents a series for the earnings of skilled relative to unskilled labor. We construct the series by combining Williamson's and Lindert's (1980, p. 307) estimates of the wage ratio for urban skilled and unskilled workers for 1870-1894 with estimates from Goldin and Katz (1999) of the ratio of clerical to manufacturing production wages for 1895-1938 and the returns to 16 versus 12 years of schooling for men for 1939-1995.[14] Despite the need to merge data from

[14]Combining several very different series into a continuous "skill premium" is necessary due to sectoral shifts in the skilled and unskilled labor forces that render some measures of skill more applicable to certain periods than others. For example, a college education appears to have become

disparate sources to form a continuous series, a U-shape still emerges, with the skill premium high in the early stages of Electrification (i.e., 1890 to 1918), and then rising rapidly during the post-1978 part of the IT epoch. We suspect that the fall in the skill premium from 1918-1924 would be less deep, and thus the overall U-shape of Figure 18 more apparent, had it not been for the rapid rise of the public higher-education system after the end of the first World War [see Goldin and Katz (1998, p. 10)].

3.3 Entry, exit, and mergers should rise

Gort (1969) argued that technological change will generate merger waves. Evidence since then has shown that mergers and takeovers play a reallocative role. Lichtenberg and Siegel (1987), McGuckin and Ngyen (1995) and Schoar (2000) find that the productivity of a target firm rises following a takeover. The trade-off between exits and acquisitions margins is studied by Jovanovic and Rousseau (2002b,c). This last pair of papers shows that firms will tend to place themselves on the merger market rather than disassemble and sell their assets at times when the value of organization capital is high. Further, reallocation of assets among firms either by merger, consolidation, or purchases of unbundled used capital are more likely to occur than purchases of new capital when firms need to make large adjustments to their capital stocks because of the fixed costs associated with entering the merger market. We believe that both of these conditions are likely to hold during times of sweeping technological change.

The U-shaped top line of Figure 19 is our estimate of the total amount of capital that has been reallocated on the U.S. stock market over the past 112 years. Its components are the stock market capitalization of entering and exiting firms divided by two, and the value of merger targets.[15] Entries and exits divided by two, given

a more important determinant of income in the postwar period than it was in earlier years. Since the Goldin and Katz observations are generally decadal, we interpolate between them to obtain an annual series for 1895 to 1995. The vertical dotted lines in Figure 18 mark the points where we need to change data sources.

[15] We identify targets for 1926-2001 using the CRSP stock files and various supplementary sources. CRSP itself identifies more than 8,000 firms that exited the database by merger between 1926 and 2001, but links less than half of them of them to acquirers. Our examination of the 2000 Edition of Financial Information Inc.'s *Directory of Obsolete Securities* and every issue of Predicasts Inc.'s *F&S Index of Corporate Change* between 1969 and 1989 uncovered the acquirers for 3,646 of these unlinked mergers, 1803 of which turned out to be CRSP firms. We also recorded all mergers from 1895 to 1930 in the manufacturing and mining sectors from the original worksheets underlying Nelson (1959) and collected information on mergers from 1885 to 1894 from the financial news section of weekly issues of the *Commercial and Financial Chronicle.* For 1890-1930, we use worksheets for the manufacturing and mining sectors that underlie Nelson (1959). The resulting target series includes the market values of exchange-listed firms in the year prior to their acquisition, and reflects 9,030 mergers. Stock market capitalizations are from our extension of CRSP backward to 1885 (see footnote 11). Before assigning a firm that no longer carries a price in our database as an "exit," we check the list of hostile takeovers from Schwert (2000) for 1975-1996 and individual issues of the *Wall Street Journal* from 1997-2001 to ensure that we record firms taken private under a hostile tender offers as mergers.

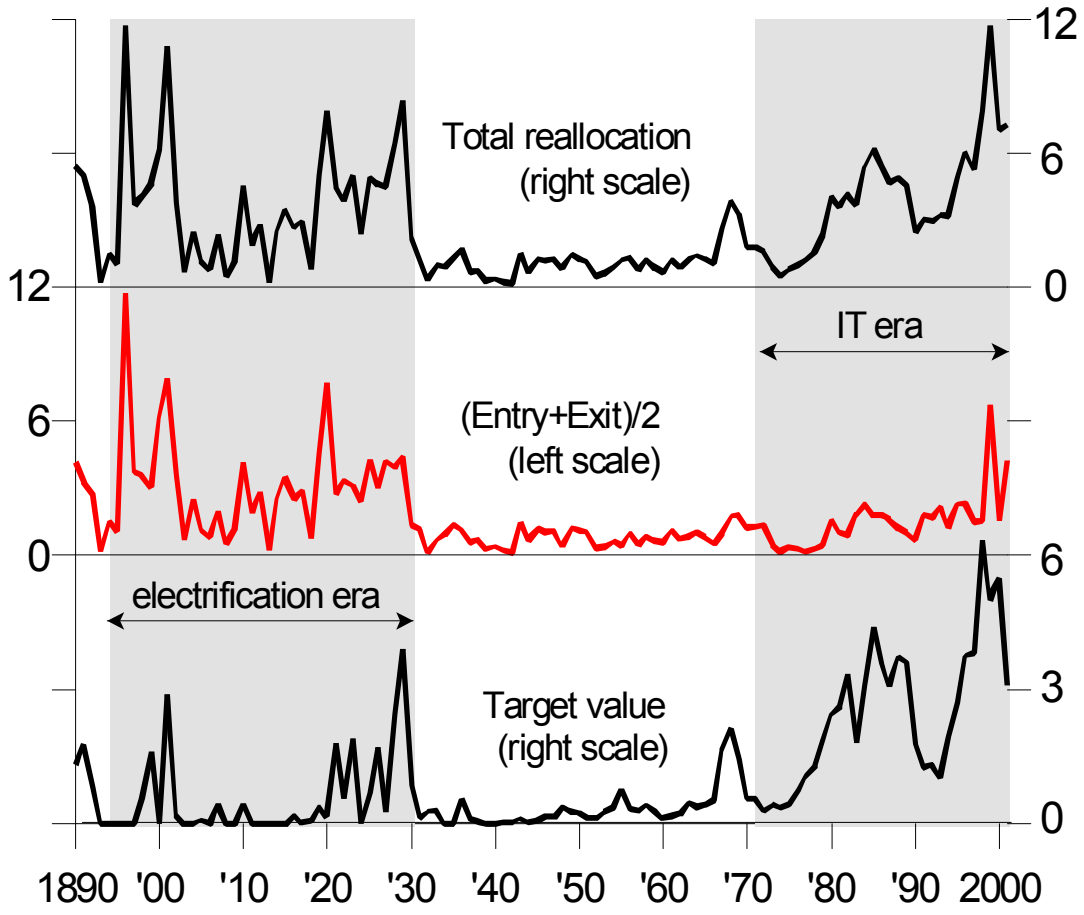


Figure 19: Reallocated capital and components as percentages of stock market value, 1890-2001.

by the center line, is a rough measure of how much capital exits from the stock market and comes back in under different ownership, or at least under a different name. The lower line is the stock-market value of merger targets. Regardless of whether reallocation occurs through mergers or through entry and exit, it is much more prevalent during the periods that we associate with Electrification and IT.

3.4 Stock prices should fall

The value of old capital should fall suddenly if the arrival of the GPT is a surprise, as in Greenwood and Jovanovic (1999) Hobijn and Jovanovic (2001), Jovanovic and Rousseau (2002a) and Laitner and Stolyarov (2002), or more gradually as in Trajtenberg and Helpman (1998). Figure 20 shows that the stock market declined in 1973-74. No such sudden drop is visible for stock prices in the early 1890's. Why not? Maybe

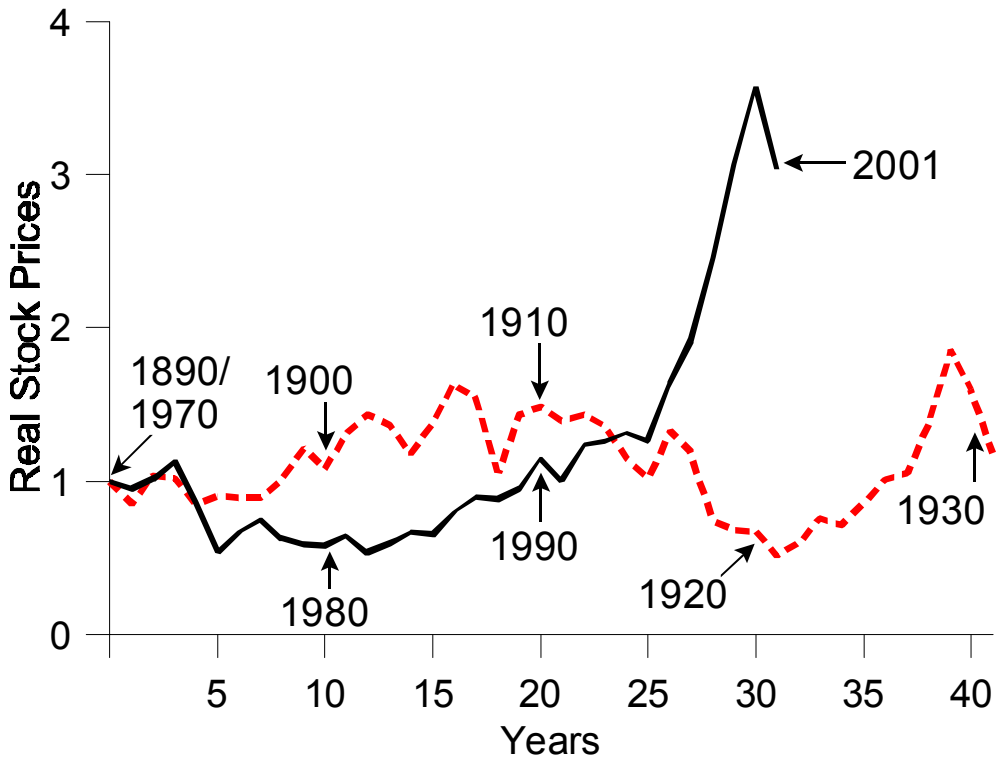


Figure 20: The real Cowles/S&P stock price index across the transition periods.

because the market was thin and unrepresentative in those days, with railway stocks absorbing a large chunk of market capitalization. More likely, the realization that the new technology would work well was more gradual and was not prompted by any single event like the completion of the Niagara Falls dam in 1894.[16]

In other words, that it does not happen for the first period could be that the events of the early 1890's were foreseen, as would be the case in Helpman and Trajtenberg (1998a) where stock prices are predicted to fall earlier and more gradually. It also could be, as Boldrin and Levine (2001) would claim, that old capital is essential in the production of new capital and that its value may not fall in quite the way that it would when capital can be produced from consumption goods alone as is the case in many growth models, including that of Jovanovic and Rousseau (2002a).

If stock-price declines were caused by the threat of IT to incumbents, this should relate especially to those sectors that later invested heavily in IT. Hobijn and Jovanovic (2001, p. 1218) confirm this using regression analysis.

[16] We obtain the composite stock price index from Wilson and Jones (2002), and deflate using the CPI.

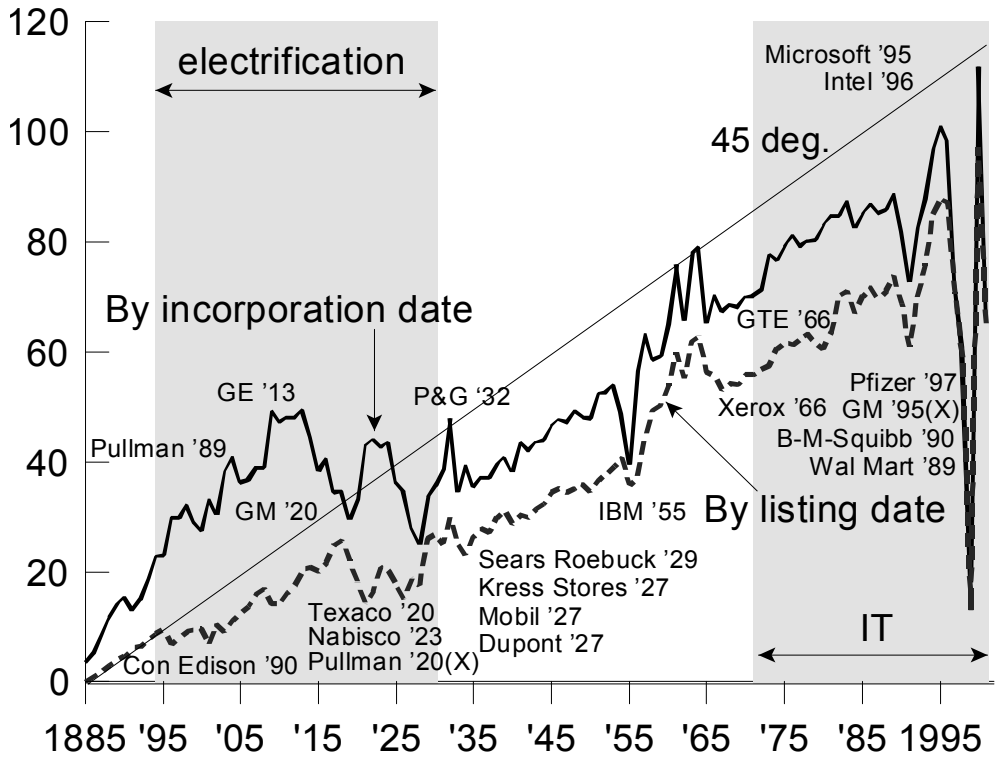


Figure 21: Average age (in years) of the largest firms whose market values sum to 5 percent of GDP.

3.5 Young firms should do better

The evidence on this hypothesis is mixed, but positive overall.

3.5.1 The age of the leadership

As a GPT takes hold, we should not only expect to see firms coming to market more quickly, but the market leaders getting younger as well. In other words, every stage in the lifetime of the firm should be shorter. This stands in contrast to Hopenhayn (1982), in which the age distribution of an industry's leadership is invariant when an industry is in a long-run stochastic equilibrium. That is, the average age of, say, the top 5% or top 10% of the firms is fixed. Some leaders hold on to their positions and this tends to make the leading group older, but others are replaced by younger firms, and this has the opposite effect. In equilibrium the two forces offset one another and the age of the leadership stays the same. Keeping the age of the leaders flat requires, in other words, constant replacement.

Figures 21 and 22 show that, overall, the age of the leaders is anything *but* flat. It sometimes rises faster than the 45^0 line, indicating that the age of the leaders is

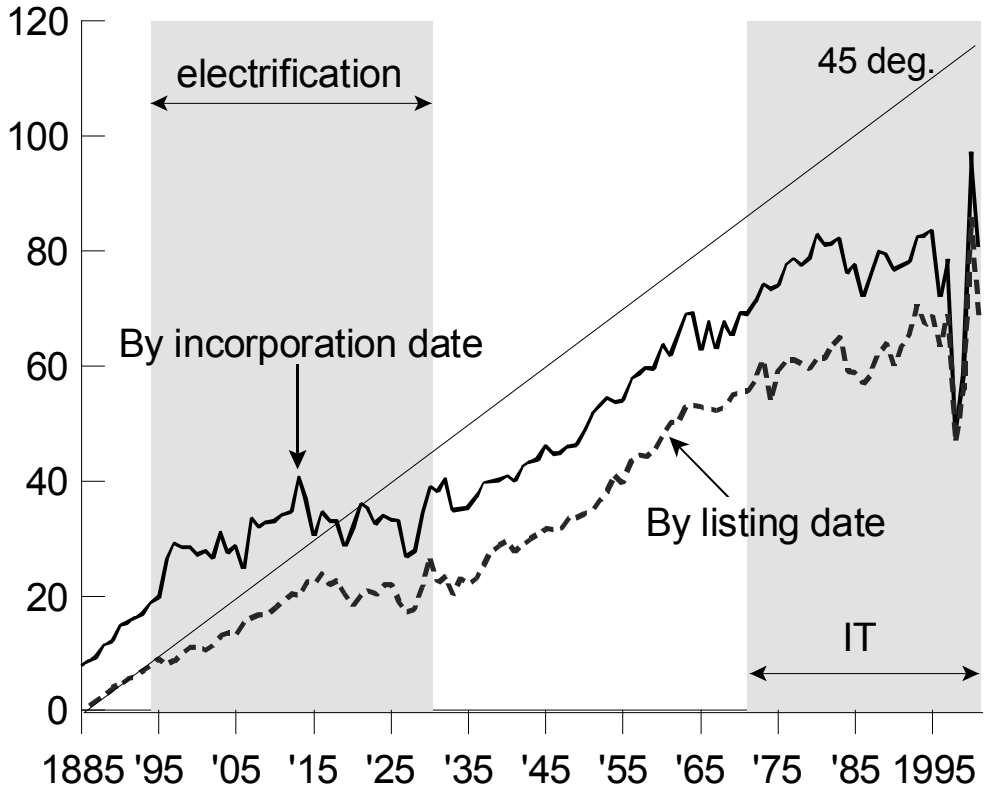


Figure 22: Average age (in years) of the largest firms whose market values sum to 10 percent of GDP.

rising faster than the passage of time. At other times it is flat or falling, indicating replacement. Both figures show, however, that during the Electricity and the IT revolutions, the lines are flat or falling, so that replacement was then high. This is best seen in Figure 22.

Figure 21 plots the value-weighted average age of the largest firms whose market values sum to 5 percent of GDP. "Age" is from incorporation and from exchange listing. We label some important entries and exits from this group (with exits denoted by "X"). Based upon years from incorporation, the leading firms were being replaced by *older* firms over the first 30 years of our sample, because the solid line is then steeper than the 45° line. In the two decades after the Great Depression the leaders held their relative positions as the 45° slope of the average age lines shows. The leaders got younger in the '90s, and their average ages now lie well below the 45° line. The shake-out of 1999-2001 comes from Microsoft's huge rise in 1999, when it was worth more than 5 percent of GDP on its own, and its rapid decline in 2000, which transferred the full 5 percent share to GE. The two firms split the 5 percent share in 2001. The slopes of regression lines in Figure 21 (estimated with constant and

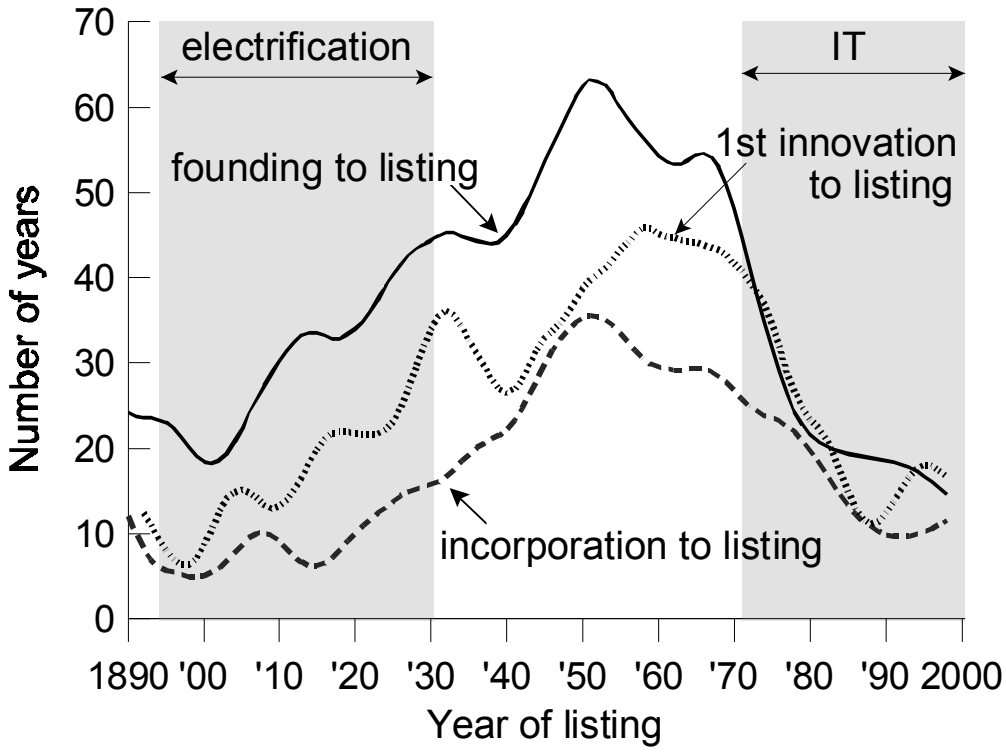


Figure 23: Waiting times to exchange listing.

time trend) are 0.62 for years since incorporation and 0.67 for years since exchange listing.

3.5.2 The age of firms at their IPO

According to its third "innovation-spawning" characteristic, when a GPT arrives it gives rise to new projects that are unusually profitable. When such projects arrive, firms will be more impatient to implement them. When it is the new firms that come upon such projects (rather than the incumbents), they will feel the pressure to list sooner. The argument is developed and tested in Jovanovic and Rousseau (2001b). We argue there that the Electricity-era and the IT-era firms came in younger because the technologies that they brought in were too productive to be kept out of the market for very long.

Figure 23 shows HP-filtered average waiting times from founding, first product or process innovation, and incorporation to exchange listing based upon individual company histories and our extended CRSP database.[17] The vertical distance between

[17]Listing years after 1925 are those for which firms enter CRSP. For 1885-1924, they are years in which prices first appear in the NYSE listings of *The Annalist, Bradstreet's*, *The Commercial*

the solid and dotted lines shows that firms often have their first innovation soon after founding, but that it then takes years, even decades, to list on a stock exchange. We interpret this delay as a period during which the firm and possibly its lenders learn about what the firm's optimal investment should be. But when the technology is highly innovative, the incentive to wait is reduced, and the firm lists earlier, which is what the evidence shows.

When firms gather less information before investing, the investments that they undertake will be riskier. One may conjecture that if new entrants waited less before investing, then incumbents also undertook projects earlier than they would normally. In all these cases, the resulting investments would be riskier than if more time were allowed to plan them. Moreover, the newness of the GPT would add further risk. On all these grounds, we expect higher interest rate differentials on the average investment.

Figure 24, which portrays the spread between interest rates on riskier and safe investments since 1885, shows that this has been for the most part the case.[18] It is important to note that we formed the series in Figure 24 by ratio-splicing three different spreads together, and that the "safe" asset after 1920 is the short-term treasury bill while for earlier years it is a long-term government bond, yet the fluctuations in this series should still reflect fluctuations in risk perceptions reasonably well, at least to the extent that term premia rather than riskiness are the main factors that lead to yield differentials among the various government securities.

During the Electrification era, spreads rose between 1894 and 1907, which is when uncertainty about the usefulness and possibilities for adoption of the new technology were greatest. Spreads fell after that as the future of electricity became more clear. In the IT-era, spreads have an upward trend from 1970 through the 1990s, and only seem to have fallen recently. This may well reflect the lag in the widespread adoption of the IT technology. The rise of the spread in 1930 and its very slow decline in the subsequent 15 years through 1945 probably has to do with the macroeconomic instability induced by the events prior and during the Great Depression, and then

and Financial Chronicle, or *The New York Times*. The 6,238 incorporation dates used to construct Figure 23 are from *Moody's Industrial Manual* (1920, 1928, 1955, 1980), Standard and Poor's *Stock Market Encyclopedia* (1981, 1988, 2000), and various editions of Standard and Poor's *Stock Reports*. The 3,827 foundings are from Dun and Bradstreet's *Million Dollar Directory* (2000), Moody's, Etna M. Kelley (1954), and individual company websites. The 482 first innovations were obtained by reading company histories in *Hoover's Online* (2000) and company websites. We linearly interpolate the series between missing points before applying the HP-filter to the time series in the figure.

[18] In Figure 24, we use the spread between the interest rates on Baa rated corporate bonds (from Moody's Investors Service) and three-month T-bills [from the FRED database for 1934-2001 and the Board of Governors (1976) for 1920-34] for the period from 1920 to the present. For 1900-20, we ratio splice the spread between the interest rate on prime commercial paper with 60-90 days until maturity [Homer and Sylla (1991, table 49, p. 358)] and the redemption yields on the U.S. government consol 2s of 1930 [Homer and Sylla (1991, table 46, p. 343] onto the Baa - T-bill spread. Finally, for 1885-99, we splice the spread between the commercial paper rate [Homer and Sylla (1991, table 44, p. 320] and the redemption yields on U.S. government refunding 4s of 1907 [Homer and Sylla (1991, table 43, p. 316] onto the previous result.

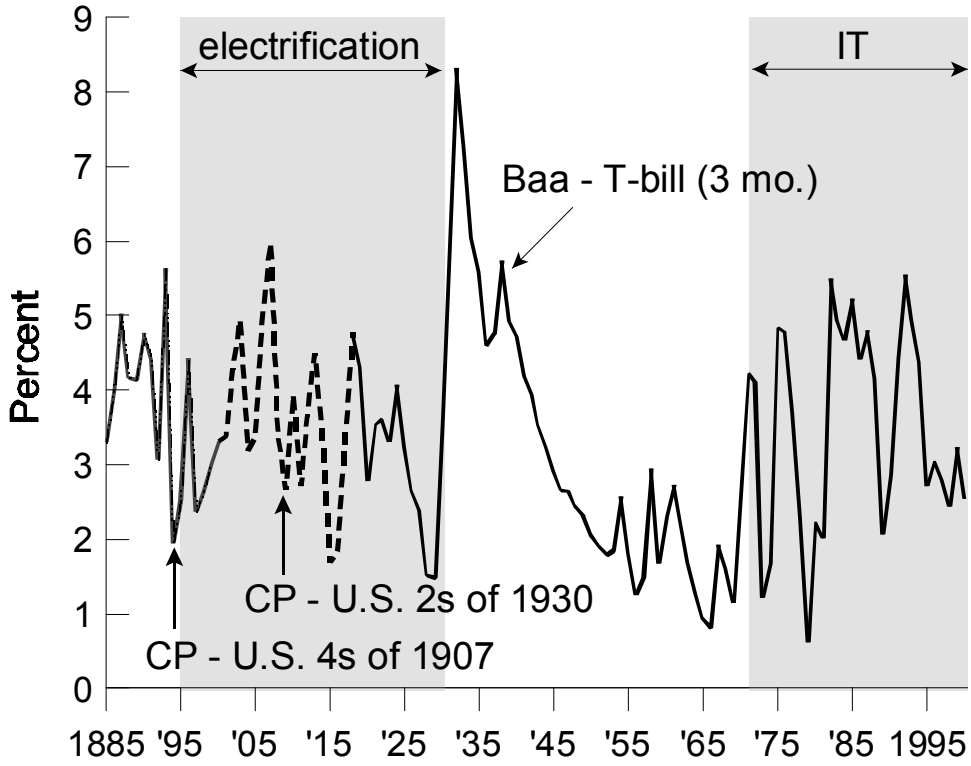


Figure 24: Nominal interest rate spreads between riskier and safer bonds, 1885-2000.

the heavy borrowing by the U.S. government to to finance the second World War, which lowered rates on T-bills.

3.5.3 The stock market performance of young vs. old after entry

Young firms are smaller. If "creative destruction" does indeed mean that old firms give way to young firms, then we should see signs of it in Figure 25, which depicts the relative appreciation of *total* market value of small versus large firms since 1885.[19] We define "small" firms as those in the lowest quintile of CRSP, and "large" firms as those in the upper quintile. The figure shows that small firms outperform large ones in the long run and that the growth premium is about 7.5 percent per year. But the two technological epochs do not show a faster rise than the other epochs, and this is puzzling. The IT era shows, in particular, that the large firms regrouped after 1983. Surprisingly, recessions do not seem to hurt the long-term prospects of small firms: The relative index rises in 10 of the 23 NBER recessions.

[19] Being a total value index, this differs from the relative stock price index that is plotted in Figure 8 of Hobijn and Jovanovic (2001). For the post-1925 period for which they overlap, the qualitative behavior of the two series is essentially the same.

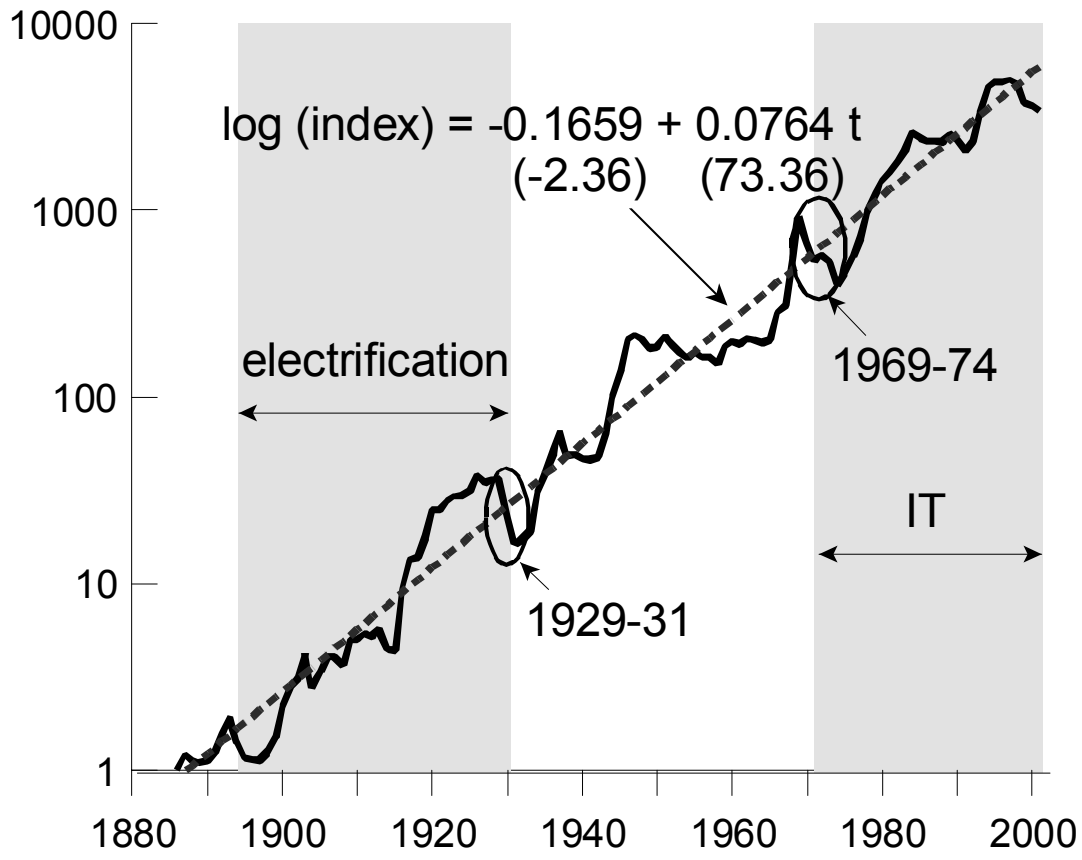


Figure 25: The relative capital appreciations of small vs. large firms.

The two periods that we wish to focus on are 1929-31 and the early 1970s. In both periods, the small-cap firms lost out relative to the large-cap firms. The first period comes at the end of the Electrification era and the relative decline of small cap firms is what one would have expected. But the early 1970's come at the beginning of a new GPT, and small-cap firms should have outperformed the large-cap firms during that time. Yet the opposite happened. It is only after 1974 that the small-cap firms start to perform better.

Regression evidence on age and stock-market performance.—If the GPT is brought in by young firms, then the capital loss imposed by the GPT's arrival should fall more heavily on old firms. To test this using data on individual firms, let

A_i = age since listing of firm i in 1970.
S_i = share (in firm i's sector) of IT capital in the capital stock in 2001.

This measures firm's exposure to the impact of the new technology of the sector. For the measure of expected performance we take the change in a firm's stock price over intervals that start in 1971 and end in 1975, 1980, 1985, 1990, and 1995. These should

reflect the market's expectation of how well the firm will handle the consequences of the GPT. We regressions take the form:

$$\ln\left(\frac{P_{i,1975}}{P_{i,1970}}\right) = c_0 + c_1 A_i + c_2 S_i - c_3 A_i S_i. :$$

We summarize the firm-level results in Table 4.

Table 4
Age and stock market performance

	Dependent variable: $\ln(P_{t+i}/P_t)$				
	1971-75	1971-80	1971-85	1971-90	1971-95
constant	-0.737	-0.143	0.152	-0.057	0.577
	(-24.3)	(-2.96)	(2.58)	(-0.59)	(6.06)
A	0.007	-0.001	-0.001	0.003	-.002
	(6.40)	(-0.46)	(-0.55)	(0.97)	(-0.51)
S	-3.497	-2.266	-1.035	-0.602	2.719
	(-7.60)	(-3.37)	(-1.20)	(-0.46)	(1.88)
$A * S$	0.047	0.043	-0.016	-0.122	-0.106
	(2.22)	(1.14)	(-0.39)	(-2.09)	(-1.76)
R^2	.089	.009	.003	.006	.012
N	2218	1814	1367	981	843

Note: The table presents coefficient estimates for the subperiods included in column headings with T-statistics in parentheses. The R^2 and number of observations (N) for each regression appear in the final two rows.

The interaction between the firm's age (A) and its exposure to (S) is negative and significant only when the period during which we measure price appreciation extends to 1990 and 1995. We would have expected this coefficient to be negative always since older firms in sectors in which IT would become important would be less able to adjust to the new technology than newer ones. Over the 1971-75, 1971-80, and 1971-85 periods, the interaction term comes in positive, but statistically significant only in the first subperiod. It thus seems that IT firms took a long time to realize gains in the market after the technology's arrival. We do not have a lot of firms with continuous price data prior to 1900, but have enough observations to at least attempt the same regression for the Electrification era. In this case, we got

$$\ln\left(\frac{P_{i,1899}}{P_{i,1894}}\right) = \underset{(1.09)}{2.111} - \underset{(-0.46)}{0.129} A_i - \underset{(-0.88)}{2.307} S_i + \underset{(0.55)}{0.213}\, A_i S_i$$

(T-stats in parenthesis) $R^2 = .015$, $N = 56$. In this very small sample, we do not see a direct effect of age on capital depreciation as the Electrification era got underway, and the interaction term is not statistically significant.

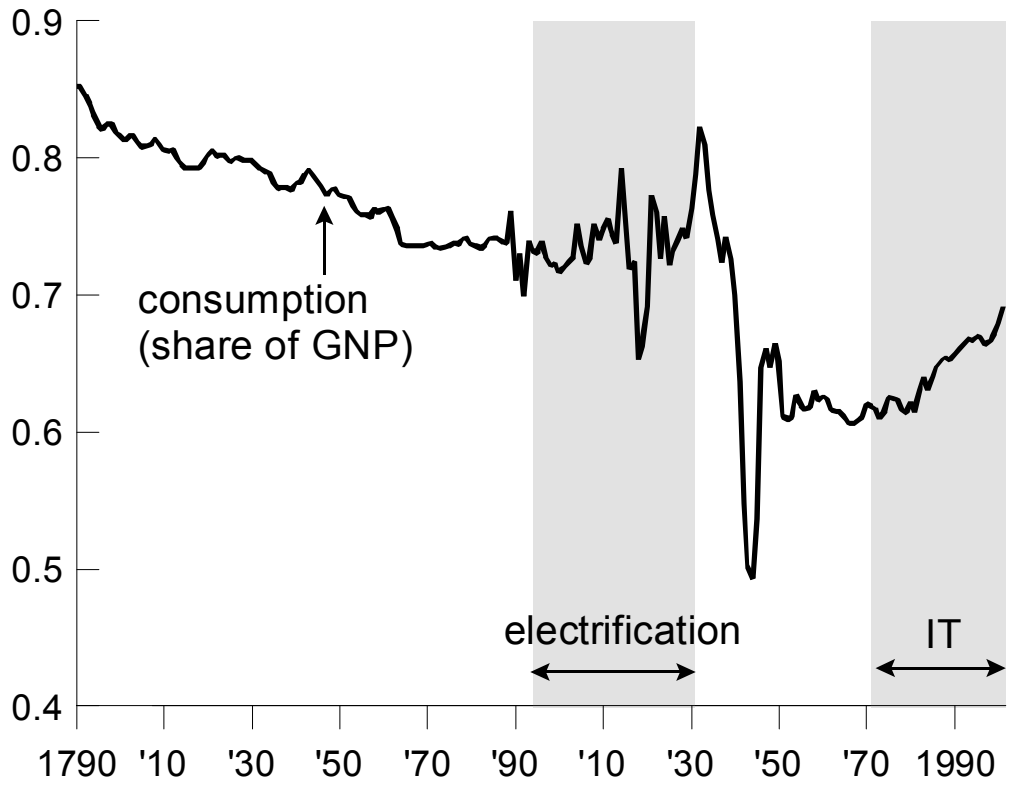


Figure 26: The ratio of consumption to income.

3.6 Interest rates and the trade deficit

First, if the productivity gains of the GDP are deferred, the consumption income ratio should rise. Second, the consumer becomes wealthier when the GPT arrives, interest rates should rise to choke off some of the increased consumption demand. Finally, some of the extra consumption comes from imports and so the trade deficit should rise. The evidence on these three point is mixed, but on balance favorable.

3.6.1 The predicted rise in the consumption-income ratio during GPT eras

Private consumption rises gradually during each GPT era, and this is set against a long-run secular trend for private consumption that is negative. Figure 26 shows the ratio of consumption to gross national product (GNP) since 1790.[20] As our GPT

[20]The series for consumption and GNP are from the Bureau of Economic Analysis (2002, table 1, pp. 123-24) for 1929-2001, Kendrick (1961, table A-IIb, cols. 4 and 11, pp. 296-97) for 1889-1929, and Berry (1988, table 9, pp. 25-26) for 1790-1889. The BEA figures are for personal consumption, but the Kendrick and Berry figures include the government sector as well. Since consumption in the

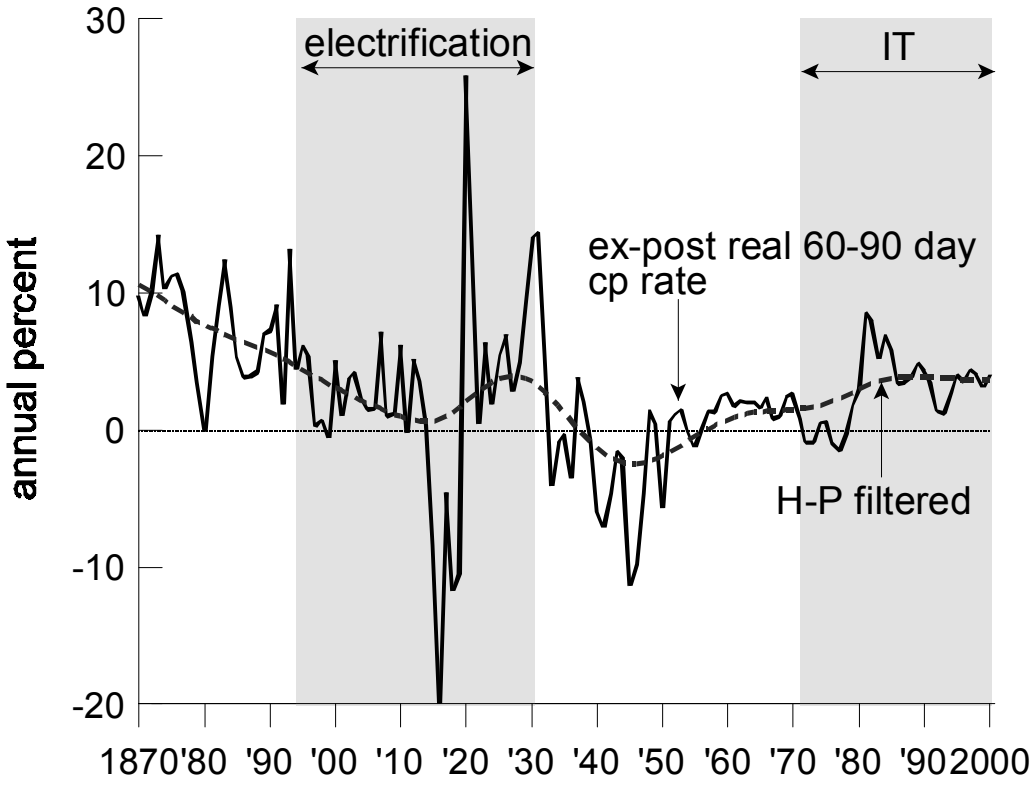


Figure 27: The ex-post real interest rate on commercial paper.

hypotheses would suggest, the arrival of Electricity in 1890 seems to mark the end of a long-term decline in the ratio that been underway for a century. And though the level of the series falls during the Great Depression and World War II, never to return to its pre-1930 levels, consumption takes another sharp upward turn near the start of the IT revolution, and has continued its rise to this day.

3.6.2 The predicted rise in interest rates

Figure 27 shows that ex-post real interest rates were about the same during the two GPT eras, and much lower in the middle 40 unshaded years of the 20th century.[21]

government sector was much smaller prior to the first World War, we suspect that the downward trend in the 19th century is a result of changing private consumption patterns rather than a reduction in the government sector's consumption.

[21] Commercial paper rates are annual averages from the FRED database for 1934-2001 and from Homer and Sylla (1991) for earlier years. We compute the ex-post return by subtracting inflation as computed by the growth of the implicit price deflator for GNP from the Bureau of Economic Analysis (2002) for 1929-2001 and Berry (1988) for earlier years.

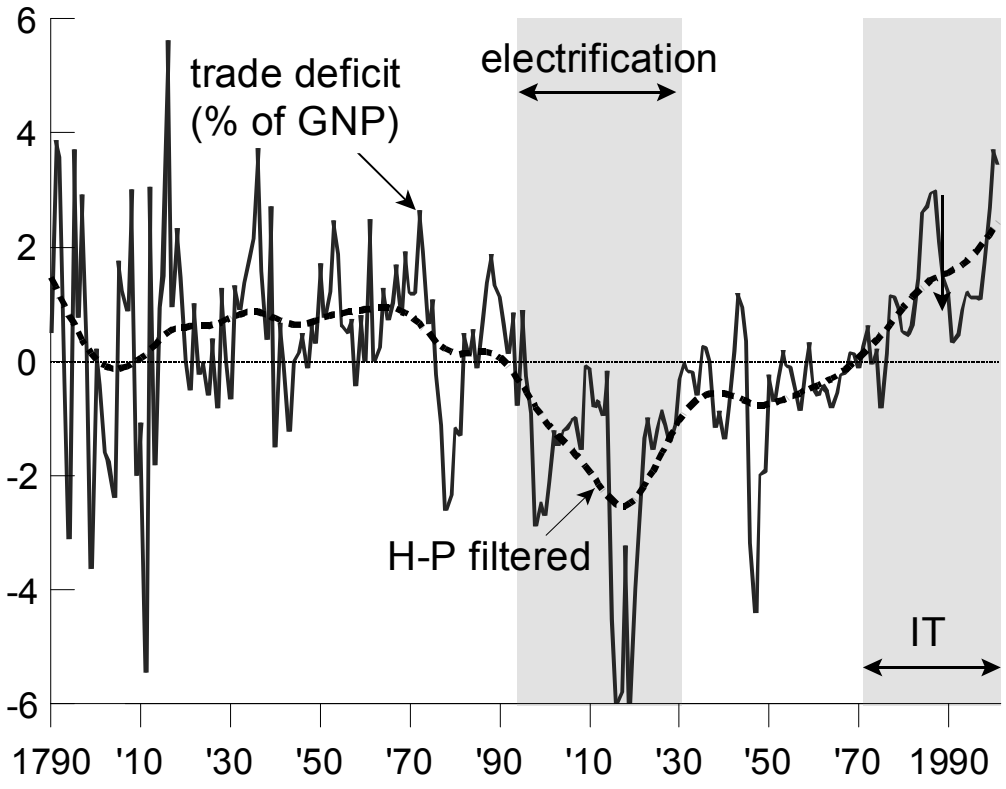


Figure 28: Trade deficit as a percent of GNP.

The dashed line is the H-P detrended series. The averages are

Era	Ex-post real interest rate
$1870 - 1893$	7.78
$1894 - 1930$	2.61
$1931 - 1970$	-0.16
$1971 - 2001$	2.94

We note, however, that the ex-post rate is so high in the first era, before 1894. If the arrival of Electricity and its impact was foreseen prior to 1994, interest rates would have risen earlier, but this probably does not explain why it is so high then. More likely, the pre-1995 era reflects the lack of financial development: The stock market was small then, and the financial market not as deep. This may have given rise to an overall negative trend in interest rates over the 130-year period as a whole.

3.6.3 The predicted rise in the trade deficit

Alternatively, a trade deficit should open up, and right away if the economy is open. Figure 28, which plots the trade deficit as a percentage of GNP since 1790 along with

an H-P trend, shows this to be the case at the start of the IT revolution, though not in the early years of Electrification.[22]

It is puzzling at first blush that the deficit did not appear during the IT era. If anything, it moves into surplus. Events in the rest of the world may explain a part of this. Europe needed to borrow heavily to finance its colonial wars and World War I, and the United States provided some of the loans. This may partly explain the surplus. The IT era is taking place while the world is largely at peace

4 Conclusion

Technological invention is uneven, and comes in bursts; that much has for a long time been clear to students of growth. Electricity and IT are, to most observers, the two most important GPTs to date, at least they seem so according to the three criteria that Bresnahan and Trajtenberg proposed. In this chapter we have analyzed how the U.S. economy reacted to the creation of these two GPTs. Having discussed in detail GPTs with reference to the Electricity and IT eras, we believe we have shown that the concept is a good way to organize how we think of technological change and its effects.

The Electricity and IT eras are similar, but they also different in important ways. Electrification was more broadly adopted, whereas IT seems to be technologically more revolutionary. The productivity slowdown is stronger in the IT era but the ongoing spread of IT and its continuing precipitous price decline are reasons for optimism about growth in the coming decades.

[22] GNP and total imports and exports of goods and services are from the Bureau of Economic Analysis (2002, table 1, pp. 123-24) for 1929-2001. For 1790-1920, imports and exports are from Bureau of Census (1975), series U-8 and U-1, p. 864, respectively), and the GNP series are from Kendrick (1961) and Berry (1988).

References

[1] Aghion, P., P. Howitt, and G. L. Violante (2002), "General purpose technology and wage inequality", manuscript (University College London, Harvard University, and CEPR Brown University).

[2] Atkeson, A., and P. J. Kehoe (1993), "Industry evolution and transition: the role of information capital", Staff Report No. 162 (Federal Reserve Bank of Minneapolis, , MN).

[3] Atkeson, A., and P. J. Kehoe (2001), "The transition to a new economy after the second industrial revolution", Working Paper No. w8676 (National Bureau of Economic Research, Cambridge, MA).

[4] Bahk, B. H., and M. Gort (1993), "Decomposing learning by doing in plants", Journal of Political Economy 101: 561-83.

[5] Bartel, A., and F. Lichtenberg (1987), "The comparative advantage of educated workers in implementing new technology", Review of Economics and Statistics 69: 1-11.

[6] Berndt, E. R., Dulberger, E. R., and Rappaport, N. J. (2000), "Price and quality of desktop and mobile personal computers: a quarter century of history," working paper (MIT Sloan School, Cambridge, MA).

[7] Berry, T. S. (1988), "Production and population since 1789: revised GNP series in constant dollars", Bostwick Paper No. 6 (The Bostwick Press, Richmond, VA).

[8] Bessen, J. (2002), "Technology adoption costs and productivity growth: the 70's as a technology transition", Review of Economic Dynamics 5: 443-69.

[9] Board of Governors of the Federal Reserve System (1976), Banking and Monetary Statistics, 1914-1941 (Board of Governors of the Federal Reserve System, Washington, DC).

[10] Boldrin, M., and D. K. Levine (2001), "Growth cycles and market crashes", Journal of Economic Theory 96: 13-39.

[11] Brady, D. S. (1966), "Price deflators for final product estimates", in: D. S. Brady, ed., Output, Employment, and Productivity in the United States After 1800 (Columbia University Press, New York) 91-116.

[12] Brainard, W. C., and J. Tobin (1968), "Pitfalls in financial model building", American Economic Review Papers and Proceedings 58: 99-122.

[13] Bresnahan, T. F., and M. Trajtenberg (1996), "General purpose technologies: 'engines of growth'?", Journal of Econometrics, Annals of Econometrics 65: 83-108.

[14] Bradstreet Co. (1885-1925, various issues), Bradstreet's (Bradstreet Co., New York).

[15] Buttrick, J. (1952), "The inside contract system", Journal of Economic History 12: 205-21.

[16] Caballero, R. J., and M. L. Hammour (1994), "The cleansing effect of recessions" American Economic Review 84: 1350-68.

[17] The Commercial and Financial Chronicle (1885-1925, various issues).

[18] Comin, D. (2002), "Comments on James Bessen's 'Technology adoption costs and productivity growth: the 70's as a technology transition'", Review of Economic Dynamics 5: 470-476.

[19] Cowles, A., and Associates (1939), Common Stock Price Indexes, Cowles Commission for Research in Economics Monograph No. 3. Second Edition (Principia Press, Bloomington, IN).

[20] Cummins, J. G., and G. L. Violante (2002), "Investment specific technical change in the United States (1947-2000): measurement and macroeconomic consequences", Review of Economic Dynamics 5: 243-84.

[21] David, P. 1991. Computer and dynamo: The modern productivity paradox in a not-too-distant mirror. In Technology and Productivity: The Challenge for Economic Policy. Paris: OECD.

[22] Devine, W. D. (1983), "From shafts to wires: historical perspectives on electrification", Journal of Economic History 43: 347-72.

[23] Dun and Bradstreet, Inc. (2000), D&B Million Dollar Directory (Dun and Bradstreet Inc., Bethlehem, PA).

[24] DuBoff, R. B. (1964), Electric Power in American Manufacturing, 1889-1958 (Ph.D. Dissertation, University of Pennsylvania).

[25] Feder, Barnaby. "Advances in Drugs, Courtesy of Computers." New York Times (August 3 1988) p. 5.

[26] Federal Reserve Bank of St. Louis (2002), FRED Database (Federal Reserve Bank of St. Louis, St. Louis, MO).

[27] Financial Information Inc. (2000), Annual Guide to Stocks: Directory of Obsolete Securities (Financial Information Inc., Jersey City, NJ).

[28] Gates, B. (1999), Business @ the Speed of Thought (Warner Books, New York).

[29] Goldin, C., and L. F. Katz (1998), "The shaping of higher education: the formative years in the United States, 1890 to 1940", Working Paper No. 6537 (National Bureau of Economic Research, Cambridge, MA).

[30] Goldin, C., and L. F. Katz (1999), "The returns to skill in the United States across the twentieth century", Working Paper No. 7126 (National Bureau of Economic Research, Cambridge, MA).

[31] Gordon, R. J. (1990), The Measurement of Durable Goods Prices (University of Chicago Press, Chicago, IL).

[32] Gordon, R. J. (2000), "Does the 'new economy' measure up to the great inventions of the past?", Journal of Economic Perspectives 14: 49-74

[33] Gort, M. (1969), "An economic disturbance theory of mergers", Quarterly Journal of Economics 94: 624-642.

[34] Gort, M., and S. Klepper (1982), "Time paths in the diffusion of product innovations", Economic Journal 92: 630-653.

[35] Greenwood, J., and B. Jovanovic (1999), "The information-technology revolution and the stock market", American Economic Review Papers and Proceedings 89: 116-122.

[36] Greenwood, J., A. Seshadri, and M. Yorukoglu (2002), "Engines of liberation", Economie d'Avant Guarde #2.

[37] Greenwood, J., Z. Hercowitz, and P. Krusell. "Long-run implications of investment-specific technological change." American Economic Review 87: 342-362.

[38] Greenwood, J., and M.Yorukoglu (1997), "1974." Carnegie-Rochester Conference Series on Public Policy 46: 49-95.

[39] Griliches, Z. (1957), "Hybrid corn: an exploration in the economics of technological change", Econometrica 25: 501-22.

[40] Griliches, Z. (1969), "Capital-skill complementarity", Review of Economics and Statistics 6: 465-468.

[41] Hoover's Inc. (2000), Hoover's Online: The Business Network (Hoover's, Inc., Austin, TX).

[42] Helpman, E., and M. Trajtenberg (1998a), "A time to sow and a time to reap: growth based on general purpose technologies", in E. Helpman, ed., General Purpose Technologies and Economic Growth (MIT Press, Cambridge, MA).

[43] Helpman, E., and M. Trajtenberg (1998b), "The diffusion of general purpose technologies", in E. Helpman, ed., General Purpose Technologies and Economic Growth (MIT Press, Cambridge, MA).

[44] Hobijn, B., and B. Jovanovic (2001), "The IT Revolution and the Stock Market: Evidence", American Economic Review 91: 1203-20.

[45] Homer, S., and R. Sylla (1991), A History of Interest Rates, 3rd Edition (Rutgers University Press, New Brunswick, NJ).

[46] Hopenhayn, H. A. (1992), "Entry, exit, and firm dynamics in long run equilibrium", Econometrica 60: 1127-1150.

[47] Hornstein, A., and P. Krusell (1976), "Can technology improvements cause productivity slowdowns?", NBER Macroeconomic Annual 1976: 209-259.

[48] Jovanovic, B., and Y. Nyarko (1996), "Learning by Doing and the Choice of Technology", Econometrica 64: 1299-1310.

[49] Jovanovic, B., and P. L. Rousseau (2001a), "Vintage organization capital", Working Paper No. 8166 (National Bureau of Economic Research, Cambridge MA).

[50] Jovanovic, B., and P. L. Rousseau (2001b), "Why wait? A century of life before IPO", American Economic Review Papers and Proceedings 91: 336-41.

[51] Jovanovic, B., and P. L. Rousseau (2002a), "Moore's law and learning-by-doing", Review of Economic Dynamics 4: 346-75.

[52] Jovanovic, B., and P. L. Rousseau (2002b), "The Q-Theory of Mergers", American Economic Review Papers and Proceedings 92: 198-204.

[53] Jovanovic, B., and P. L. Rousseau (2002c), "Mergers as reallocation", Working Paper No. 9277 (National Bureau of Economic Research, Cambridge MA).

[54] Kelley, E. M. (1954), The Business Founding Date Directory (Morgan and Morgan, Scarsdale, NY).

[55] Kendrick, J. (1961), Productivity Trends in the United States (Princeton University Press, Princeton, NJ).

[56] Kortum, S., and J. Lerner (1998), "Stronger protection or technological revolution: what is behind the recent surge in patenting?" Carnegie-Rochester Conference Series on Public Policy 48: 247-304.

[57] Krusell, P., L. E. Ohanian, J. V. Rios-Rull, and G. L. Violante (2000), "Capital-skill complementarity and inequality: a macroeconomic analysis", Econometrica 68: 1029-53.

[58] Kuznets, S. (1961a), Capital in the American Economy: Its Formation and Financing (Princeton University Press, Princeton, NJ).

[59] Kuznets, S. (1961b), "Annual estimates, 1869-1955" manuscript (Johns Hopkins University, Baltimore, MD).

[60] Laitner, J., and D. Stolyarov (2002), "Technological change and the stock market", manuscript (University of Michigan, Ann Arbor, MI).

[61] Lerner, J. (2001), "150 years of patent protection", Working Paper (National Bureau of Economic Research, Cambridge, MA).

[62] Matsusaka, J. (1996), "Did tough antitrust enforcement cause the diversification of American corporations?", Journal of Financial and Quantitative Analysis 31: 283-94.

[63] McGowan, J. (1971), "International comparisons of merger activity", Journal of Law and Economics 14: 233-50.

[64] Moody's Investors Service (1920, 1929, 1931, 1941, 1951, 1956, 1961), Moody's Industrial Manual. (Moody's Investors Service, New York.

[65] Lichtenberg, F., and D. Siegel (1987), "Productivity and changes in ownership of manufacturing plants", Brookings Papers on Economic Activity, Special Issue On Microeconomics, No. 3: 643-673.

[66] McGuckin, R., and S. Ngyen (1995), "On productivity and plant ownership change: new evidence form the longitudinal research database", Rand Journal of Economics 26: 257-76.

[67] Nelson, D. (1995), Managers and Workers. 2nd Edition (University of Wisconsin Press, Madison, WI).

[68] Nelson, R. L. (1959), Merger Movements in American Industry, 1895-1956 (Princeton University Press, Princeton, NJ).

[69] Nelson, R., and E. Phelps (1966), "Investment in humans, technological diffusion, and economic growth", American Economic Review 56: 69-79.

[70] The New York Times Co. (1897-1928, various issues), The New York Times (The New York Times Co., New York).

[71] The New York Times Co. (1913-1925, various issues), The Annalist: A Magazine of Finance, Commerce, and Economics (The New York Times Co., New York).

[72] Predicasts, Inc. (1969-1992), Predicasts F&S Index of Corporate Change (Predicasts Inc, Cleveland, OH).

[73] Rousseau, P. L. (1999), "Share liquidity and industrial growth in an emerging market: the case of New England, 1854-1897", Historical Working Paper No. 103 (National Bureau of Economic Research, Cambridge, MA).

[74] Schoar, A. (2000), "Effects of corporate diversification on productivity", manuscript (MIT Sloan School, Cambridge, MA).

[75] Singh, A. (1975), "Take-overs, economic natural selection, and the theory of the firm: evidence from the postwar United Kingdom experience", Economic Journal 85: 497-515.

[76] Standard and Poor's Corporation (2002), Compustat database, (Standard and Poor's Corporation, New York).

[77] Standard and Poor's Corporation (1981, 1988, 2000), Stock Market Encyclopedia (Standard and Poor's Corporation, New York).

[78] Standard and Poor's Corporation (various years), Stock Reports (Standard and Poor's Corporation, New York).

[79] United States Bureau of the Census, Department of Commerce (1975), Historical Statistics of the United States, Colonial Times to 1970 (Government Printing Office, Washington DC).

[80] United States Bureau of Economic Analysis (2002), Survey of Current Business (Government Printing Office, Washington, DC).

[81] University of Chicago Center for Research on Securities Prices (2002), CRSP Database (University of Chicago Center for Research on Securities Prices, Chicago, IL).

[82] Wiggins, S. N., and D. G. Raboy (1996), "Price premia to name brands: an empirical analysis", Journal of Industrial Economics 44: 377-88.

[83] Williamson, J. G., and P. H. Lindert (1980), American Inequality: A Macroeconomic History (Academic Press, New York).

[84] Wilson, J. W., and C. P. Jones (2002), "An analysis of the S&P 500 index and Cowles's extensions: price indexes and stock returns, 1870-1999", Journal of Business 75: 505-33.

The Effects of Technical Change on Labor Market Inequalities*

Andreas Hornstein† Per Krusell‡ Giovanni L. Violante§

First Draft: September 24, 2003
This Draft: December 6, 2004

Abstract

In this chapter we inspect economic mechanisms through which technological progress shapes the degree of inequality among workers in the labor market. A key focus is on the rise of U.S. wage inequality over the past 30 years. However, we also pay attention to how Europe did not experience changes in wage inequality but instead saw a sharp increase in unemployment and an increased labor share of income, variables that remained stable in the U.S. We hypothesize that these changes in labor market inequalities can be be accounted for by the wave of capital-embodied technological change, which we also document. We propose a variety of mechanisms based on how technology increases the returns to education, ability, experience, and "luck" in the labor market. We also discuss how the wage distribution may have been indirectly influenced by technical change through changes in certain aspects of the organization of work, such as the hierarchical structure of firms, the extent of unionization, and the degree of centralization of bargaining. To account for the U.S.-Europe differences, we use a theory based on institutional differences between the United States and Europe, along with a common acceleration of technical change. Finally, we briefly comment on the implications of labor market inequalities for welfare and for economic policy.

Keywords: Inequality, Institutions, Labor Market, Skills, Technological Change.

JEL Classification: D3, J3, O3.

*Prepared for the *Handbook of Economic Growth* (Philippe Aghion and Steven Durlauf, Editors). We are grateful to Philippe Aghion for his suggestions on how to improve an early draft. We thank Stephan Fahr, Giammario Impullitti, Matthew Lindquist, and John Weinberg for comments, Hubert Janicki for research assistance and Eva Nagypal and Bruce Weinberg for providing their data. Krusell thanks the NSF for research support. Violante thanks the CV Starr Center for research support. Any opinions expressed are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. For correspondence, our e-mail addresses are: andreas.hornstein@rich.frb.org, pkrusell@princeton.edu, and gianluca.violante@nyu.edu.

†Federal Reserve Bank of Richmond.

‡Princeton University, Institute for International Economic Studies, CAERP, CEPR, and NBER.

§New York University and CEPR.

1 Introduction

In this chapter we discuss the recent three decades of data on technology, productivity, and labor market outcomes. In particular, we explore the hypothesis that technological change has affected the labor market in various ways. We argue that (i) there is ample evidence indicating significant capital-embodied and/or skill-biased technological change and that (ii) this kind of technological change would plausibly lead to many of the transformations in the labor markets that we have observed. On the one hand, we are interested in possible implications of non-neutral technological change—of the kind we think we have experienced—on variables like wage inequality, unemployment, labor share, and unionization. On the other hand, we explore the possibility that the labor market can be used as an additional source of evidence of non-neutral technological change, a testing ground of sorts.

The past 30 years are particularly informative because they have contained rather important trend changes in several variables. We have seen a productivity slowdown common to all industrialized countries and common to almost all industries, together with continuing structural change away from manufacturing and toward services. An exception to this widespread productivity slowdown was the fast and accelerating productivity growth of in the industries producing investment goods, in particular those producing equipment. Only very recently has there been a more widespread acceleration of productivity growth. Of course, in this context we are arguably in the midst of an "Information Technology Revolution." We also discuss evidence of changes in the workplace—in how production within firms is organized—possibly reflecting underlying changes in technology.

In the labor market, we have seen a sharp increase in wage inequality in the United States contrasting a roughly flat development in Europe, whereas we have witnessed a strong increase in European unemployment and no trend in U.S. unemployment.[1] The organization of labor markets seems to have changed too: for example, unions have lost prevalence during this period, and to the extent there have been unions, centralized bargaining has been

[1] Although the word inequality literally would suggest a zero-one classification—either there is inequality or there is equality—we will use the term loosely to reflect some measure of dispersion. That is, we will attach quantifiers such as "more" or "less" to the word.

replaced by decentralized bargaining in many sectors. Are all these developments consistent with basic economic theory and a short list of underlying technological driving forces? We argue that they are. To make our argument more convincing, we also put the past three decades in a historical perspective, going back as far as the early 20th century with data on technological change and the skill premium.

One distinctive feature of this literature is that the many different ideas have been presented in a wide variety of theoretical frameworks ranging from the neoclassical Ramsey-Cass-Koopmans growth model to the Schumpeterian endogenous growth model; from the traditional McCall search model to the Lucas-Prescott island economy; from the Mortensen-Pissarides matching model to the competitive directed search framework; and from the Bewley-Aiyagari incomplete-markets model to Arrow-Debreu economies with limited enforcement. We think two main reasons exist for the lack of a unified framework of analysis. First, this field of research is still relatively young; second, departing from the competitive model in studying labor markets is fairly natural, and many alternative frameworks exist that incorporate frictions. The main drawback of the lack of a unifying framework—we will repeat it often in the chapter—is that making structurally based quantitative comparisons between different mechanisms is difficult.

To us, these heterogeneous approaches pose a formidable challenge in the exposition. Our solution has been to give priority to presenting a range of ideas, using a variety of theoretical setups, rather than to discuss in great detail a few more specific frameworks. This approach has necessitated a summarization of some rather rich models in a few key equations, which misses some of the elegance and richness of the original frameworks. We hope, however, that our spanning a wide spectrum of ideas and macroeconomic effects of technological change helps paint a picture that is broader and that, at least in an impressionistic way, suggests that the main underlying hypothesis we are proposing is quite reasonable.

The presentation of the ideas in this chapter is organized into four parts. In the first part, Section 2, we review the main trends in the data on technological change and labor market inequalities. We then cover two kinds of theories that could account for the data.

In the second part of the chapter we cover "neoclassical" theory (i.e., models where wages directly reflect marginal productivity). We view the firm as hiring labor of different skill levels in a competitive and frictionless labor market. Wages, thus, will be influenced by technology in a very direct way. Similarly, the returns to education, ability, and experience, which we discuss in detail in Sections 3 and 4, respectively, will be directly tied to changes in technology. Therefore, within these kinds of theories, the shape of the production function of the firm is crucial. We then move beyond the production function of the firm or, rather, we attempt to go inside it. In particular, Section 5 explores the possibility that the organization of the workforce also has changed within firms. These transformations, of which there is some documentation, are arguably also a result of the kind of technological change we look at in this chapter. We point, in particular, to a recent literature that explores how firms are organized and how the IT revolution, by inducing organizational changes in the firm, had a substantial impact on wage inequality.

The second class of theories we cover, in the third part of the chapter, relies more on frictions in the labor market and deals more directly with how this market is organized. Here, technological change can still directly influence wages but there are new channels. For one, wages may not only reflect marginal productivity. Moreover, now unemployment is more in focus and is a function of technology, and since unemployment—through workers' outside option—may also feed back into wages, the picture becomes yet more complex. In the context of how wages are set, we furthermore argue in Section 6 that the importance of unions and their modus operandi are influenced by technology and, more generally, that labor income as a share of total income may respond to technological change in the presence of unions. An important point that we make in Section 7 is that "luck" can be a key part of wage outcomes for individuals active in a labor market with frictions, such as the search/matching frameworks, and that the "return to luck" can be greatly affected by technology as well. Finally, government participation in labor markets—labor-market "institutions," in the form of unemployment benefits, firing costs, and so on—likely interacts with technology in determining outcomes, and Section 8 completes the third part of this chapter

by analyzing the interaction between technological shocks and labor market institutions in the context of the comparison between the United States and Europe.

The fourth and final part of the chapter asks the "So what?" question: given the significant transformations observed, is a government policy change called for? Our discussion here is very brief. It mainly points out that a basic element underlying any decisions on policy, namely, what the welfare outcomes of the changes in wages, unemployment, and so on, are for different groups in society, is studied only partially in the literature so far. Studies of changes in expected lifetime income of different groups exist, but it is reasonable to assume that risk matters too, especially with trend changes as large as those observed (at least to the extent they are hard to foresee and insure). In Section 9, we therefore cover some examples of more full-fledged attempts to look at the distribution of consumption and welfare outcomes of the changes in technology/labor market outcomes. Finally, Section 10 concludes the chapter.

2 A Look at the Facts

Before modelling the economic forces that connect changes in technology to labor market outcomes, it is useful to begin by summarizing how labor market inequalities and the aggregate technological environment evolved over the past three decades.

2.1 Labor Market Inequalities

In the late 1980s and early 1990s, an extensive body of empirical work has started to systematically document the changes in the U.S. wage structure over the past three decades. Levy and Murnane (1992) give the first overview of an already developed empirical literature. To date, Katz and Autor (1999) and, more recently, Eckstein and Nagypal (2004), offer the most exhaustive description of the facts. In between, numerous other papers have contributed significantly to our understanding of the data on wage inequality.[2]

[2]We refer the reader to the bibliographic lists in Levy and Murnane (1992), Katz and Autor (1999), and Eckstein and Nagypal (2004) for more details.

The typical data source used in the empirical work on the subject is the sequence of yearly cross-sections in the March *Current Population Survey (CPS)*. The other important data source is the longitudinal *Panel Study of Income Dynamics (PSID)*. In this section, we limit ourselves to stating the main facts and briefly commenting on them, omitting the details on the data sets, the sample selection, and the calculations that can be found in the original references. Unless otherwise stated, the data refer to a sample of male workers with strong attachment to the labor force, i.e., full-time, full-year workers.[3]

Observation 1 Wage inequality in the United States is today at its historical peak over the post World War II period. However, early in the century it was even larger. The returns to college and high school fell precipitously in the first half of the century and then rose again until now (Goldin and Katz, 1999).

In other words, the time series for inequality over the past 100 years is "U-shaped." Although the bulk of this chapter is devoted to interpreting the dynamics of the wage structure over the past three decades, it is useful to put the evidence in a historical perspective to appreciate that the high current level of inequality is not a unique episode in U.S. history. The rest of the facts characterize the evolution of inequality since the mid 1960s.[4]

Observation 2 Wage inequality increased steadily in the United States starting from the early 1970s. The 90-10 weekly wage ratio rose by 35 percent for both males and females in the period 1965-1995: from 1.20 to 1.55 for males, and from 1.05 to 1.40 for females. The increase in inequality took place everywhere in the wage distribution: both the 90-50 differential and the 50-10 differential rose by comparable amounts (Katz and Autor, 1999).

[3]Eckstein and Nagypal (2004) systematically document all the facts for males and females separately. Typically, measures of inequality in the literature refer to hourly or weekly wages, that is, they isolate the evolution of the "price" of certain labor market skills. The use of hourly or weekly wages then avoids the contamination of the data with endogenous labor supply decisions that, for example, is present in annual earnings.

[4]In Section 5, we return briefly to this historical pattern. In passing, we note that the data seems at odds with the so-called "Kuznets Hypothesis," i.e., the conjecture that income inequality first increases and then decreases as economies grow.

Qualitatively, the rise in inequality is present independently of the measure of dispersion and of the definition of labor income. For example, the standard deviation of log wages for males rose from 0.47 in 1965 to 0.62 in 1995, the Gini coefficient jumped from 0.25 to 0.34 (Katz and Autor, 1999), and the mean-median ratio rose from 1.00 to 1.18 over the same period (Eckstein and Nagypal, 2004). Inequality of annual earnings increased even more.[5]

Observation 3 The average and median wage have remained constant in real terms since the mid-1970s. Real wages in the bottom of the wage distribution have fallen substantially. For example, the 10th wage percentile for males declined by 30 percent in real terms from 1970 to 1990 (Acemoglu, 2002a).[6] On the contrary, salaries in the very top of the wage distribution have grown rapidly. In 1970, the workers in the top 1 percent of the wage distribution held 5 percent of the U.S. wage bill, whereas in 1998 they received over 10 percent (Piketty and Saez, 2003).

A large part of the absolute increase of top range salaries is associated with the surge in CEO compensation. Piketty and Saez (2003) document that in 1970 the pay of the top 100 CEOs in the United States was about 40 times higher than the average salary. By 2000 those CEOs earned almost 1,000 times the average salary.

We now list a set of facts on the evolution of *between-group* inequality, i.e. inequality between groups of workers classified by observable characteristics (e.g., gender, race, education, experience, occupation). For this purpose, it is useful to write wages w_{it} using the Mincerian representation

$$\ln w_{i,t} = X'_{i,t} p_t + \omega_{it}, \tag{1}$$

where X_{it} is a vector measuring the set of observable features of individual i at time t, p_t can be interpreted as a vector of prices for each characteristic in X, and ω_{it} is the residual unobserved component.

[5]The reason is, perhaps surprisingly, not a rise in the cross-sectional variance of hours worked, but rather a substantial increase in the wage-hours correlation over the past 30 years. See Heathcote et al. (2003) for an account of these facts.

[6]Note, however, that the wages of the 10th wage percentile have started to increase again since the late 1990s (Eckstein and Nagypal, 2004).

Observation 4 The returns to education increased slightly from 1950 to 1970, fell in the 1970s, increased sharply in the 1980s, and continued to increase, although at a slower pace, in the 1990s. For example, the college wage premium—defined as the ratio between the average weekly wage of college graduates (at least 16 years of schooling) and that of workers with at most a high school diploma (at most 12 years of schooling)—was 1.45 in 1965, 1.35 in 1975, 1.50 in 1985, and 1.70 in 1995 (Eckstein and Nagypal, 2004). If one estimates the coefficient on educational dummies in a standard Mincerian wage regression like (1), the finding is similar: the annual return to a college degree (relative to a high-school degree) was 33 percent in the 1980s and over 50 percent in the 1990s (Eckstein and Nagypal 2004).

We plot the college wage premium over the period 1963-2002 in Figure 1 (top panel).[7] Interestingly, if one slices up the college-educated group more finely into workers with post-college degrees and workers with college degree only, the rise in the skill premium is still very apparent. The return to post-college education relative to college education doubled from 1970 to 1990 (Eckstein and Nagypal 2004).

FIGURE 1

Observation 5 The returns to professional and white-collar occupations relative to blue-collar occupations display dynamics and magnitudes similar to the data stratified by education. For example, the professional-blue collar premium rose by 20 percent from 1970 to 1995 (Eckstein and Nagypal 2004).

Occupation is an interesting dimension of the wage structure that, until recently, received very little attention. For example, the "returns to occupation" appear large and significant, over and beyond returns to education. We discuss the theories of wage inequality that stress the changes in occupational structure in Section 7.

[7] Authors differ in their treatment of workers who have attended college for some years, but did not obtain a college degree. In Figure 1 (top panel), we have followed the bulk of the literature and assigned half of them to the numerator and half of them to the denominator (e.g., Autor et al. 1998).

Observation 6 The returns to experience increased in the 1970s and the 1980s and leveled off in the 1990s. For example, the ratio of weekly wages between workers with 25 years of experience and workers with 5 years of experience rose from 1.3 in 1970 to 1.5 in 1995 (Katz and Autor, 1999). An analysis by education group shows that the experience premium rose sharply for high-school graduates but remained roughly constant for college graduates (Weinberg, 2003b).

It is worth emphasizing, that although entry of the baby-boomers into the labor market in the early 1970s had a significant impact on the experience premium, the dynamics described above are robust to this and other demographic effects. See for example, Juhn et al. (1993).[8]

Observation 7 Inequality across race and gender declined since 1970. The black-white race differential, for workers of comparable experience, fell from 35 percent in 1965 to 20 percent in 1990 (Murphy and Welch, 1992). The female-male wage gap fell from 45 percent in 1970 to 30 percent in 1995 (Katz and Autor, 1999).

We plot of the gender wage gap over the period 1963-2002 in Figure 1 (bottom panel). A unifying theory of the changes in the wage structure based on technological change should have something to say about gender as well as race. Admittedly, these two dimensions of inequality have been largely neglected by the literature. We return briefly to the gender gap in Section 4.

Observation 8 The composition of the working population changed dramatically over the past 40 years: in the period 1970-2000, women's labor force participation rate rose from 49 percent to 73 percent; college graduates rose from 15 to 30 percent of the male labor force and from 11 to 30 percent of the female labor force; professionals soared from 24 to 33 percent of the male labor force and from 8 to 28 percent of the female labor force (Eckstein and Nagypal, 2004).

[8]More recently, however, Card and Lemieux (2001) have argued in support of some "vintage effects" in the return to education. In particular, they argue that the college-high school premium is somewhat larger among the most recent cohorts of young workers entering the labor market.

We plot the relative supply of skilled workers and female workers over the time period 1963-2002 in Figure 1 (top and bottom panel, respectively).[9]

In terms of equation (1) , one can define the between-group component of wage inequality as the cross-sectional variance of $X'_{it}p_t$, and the within-group component as the variance of the residual ω_{it}. The fraction accounted for by observable characteristics, in turn, can be decomposed into what is caused by a change in the dispersion in the quantities of observable characteristics (X_{it}), for given vector of prices, and what is due to a change in the prices associated to each observable characteristic (p_t), for a given distribution of quantities.

Observation 9 Overall, changes in quantities and prices of observable characteristics (gender, race, education, experience) explain about 40 percent of the increase in the variance of log wages from 1963 to 1995. The price component is by far larger than the quantity component. Increasing *within-group inequality*, i.e., wage dispersion within cells of "observationally equivalent" workers accounts for the residual 60 percent of the total increase. With respect to the timing, the rise in within-group inequality seems to anticipate that of the college premium by roughly a decade (Juhn et al. 1993).[10]

One can specify further the structure of the residual ω_{it} of equation (1), for example as

$$\omega_{it} = \phi_t \alpha_i + \varepsilon_{it},$$

where α_i is the permanent part of unobservable skills (e.g., "innate ability"), ϕ_t is its time-varying price, and ε_{it} is the stochastic component due to earnings shocks whose variance is also allowed to change over time. If one is prepared to assume that the distribution of innate ability in the population is invariant, then with the help of panel data one can separate the rise in the return to ability from the increase in the volatility of transitory earnings shocks.

Observation 10 Around one-half of the rise in residual earnings inequality is explained by the permanent components (e.g., a higher return to ability), with the rest accounted

[9]Skilled and unskilled labor are defined as in footnote 7.

[10]Notice that, typically, occupation is excluded from these regressions. Including occupation would reduce the fraction of unexplained wage variance.

for by transitory earnings shocks (Gottschalk and Moffitt 1994).[11]

Interestingly, the rise in the transitory component is not due to higher job instability or labor mobility (Neumark, 2000), but rather to more volatile wage dynamics, in particular faster wage growth on the job and more severe wage losses upon displacement (Violante 2002).

In Table 1 we report some key numbers on unemployment, wage inequality, and labor income shares for several OECD countries at five-year intervals from 1965 to 1995. We are particularly interested in the comparison between the United States and *continental* European countries (averaged in the row labeled Europe Average). For completeness, we include data for the United Kingdom and Canada, whose behavior falls somewhere between that of the United States and continental Europe.

TABLE 1

Observation 11 The time pattern of wage inequality over the past 30 years differs substantially across countries. The U.K. economy had a rise in wage inequality similar to that in the U.S. economy, except for the fact that the average real wage in the United Kingdom has kept growing (Machin 1996). Continental European countries had virtually no change in wage inequality, whereas over the same period they had large increases in their unemployment rates (roughly, all due to longer unemployment durations) and a sharp fall in the labor income share in GDP. On the contrary, in the United States both the unemployment rate and the labor share have remained relatively constant (Blanchard and Wolfers, 2000).

In 1965 the unemployment rate in virtually every European country was lower than in the United States. Thirty years later, the opposite was true: the U.S. unemployment rate rose only by 1.7 percent from 1965 to 1995, whereas the average unemployment rate increase of European countries was 8.4 percent.[12]

[11]Note that a rise in the return on ability does generate an increase in cross-sectional variation of wages because it multiplies individual ability in the log-wage Mincerian equation.

[12]Notice, however, that in the United States non-participation of the low-skilled males rose from 7 percent to 12 percent from the early 1970s to the late 1990s (Juhn, 1992 and Murphy and Topel, 1997).

The labor income share has declined only marginally in the United States—by 1.5 percentage points from 1965 to 1995—while on average it fell by almost 6 points in Europe. Wage inequality, measured by the percentage differential between the ninth and the first earnings deciles for male workers, rose only slightly in Europe by 4 percent in the period from 1980 to 1995, and it even declined in some countries (Belgium, Germany, and Norway). Recall that, over the same period, earnings inequality surged in the United States: the OECD data show a rise of almost 30 percent, close to the numbers we reported earlier in this section.

Interestingly, the European averages hide much less cross-country variation than one would expect, given the raw nature of the comparison. For example, in 11 out of the 14 continental European countries, the increase in the unemployment rate has been larger than 6 percentage points, and in 9 countries the decline in the labor share has been greater than 5 percentage points.

Recently, Rogerson (2004) has argued that if one focuses on *employment* rate differences between the United States and Europe rather than on unemployment rate differences, a new set of insights emerges from the data. Employment rates in the United States start to increase relative to European employment rates twenty years before the divergence in unemployment rates. Moreover, the increase in European unemployment rates is correlated with the decline of European manufacturing employment.

2.2 Technological Change

The standard measure of aggregate technological change, total factor productivity (TFP), does not distinguish between the different ways in which technology grows. First, technology growth may differ across final-output sectors and second, it may have different effects on the productivity of different input factors. The recent experience of developed countries, however, seems to suggest that in the past 30 years technological change has originated in particular sectors of the economy and has favored particular inputs of production.

Arguably, the advent of microelectronics (i.e., microchips and semiconductors) induced a sequence of innovations in information and communication technologies with two features.

First, *sector-specific* productivity (SSP) growth substantially increased the productivity of the sector that produces new capital equipment, making the use of capital in production relatively less expensive. Second, *factor-specific* productivity (FSP) growth favored skilled and educated labor disproportionately. In other words, the recent technological revolution has affected the production structure in a rather asymmetric way.

Our assessment of the importance of SSP and FSP changes relies heavily on observed movements in relative prices. For SSP change, we rely on the substantial decline of the price of equipment capital relative to the price of consumption goods, a process that does not show any sign of slowing down. On the contrary, it shows an acceleration in recent years. For FSP change, we rely on the substantial increase in the wage of highly educated workers relative to less educated workers, the skill premium.

We first review the Solow growth accounting methodology for TFP within the context of the one-sector neoclassical growth model and then introduce SSP accounting and how it applies to the idea of capital-embodied technical change.[13] Next, we discuss how an acceleration of capital-embodied technical change might relate to the much-discussed TFP growth slowdown in the '70s and '80s; here, we discuss the possible relevance of the concept of General Purpose Technologies (GPTs). Finally, we explain the mapping between relative wages and FSP changes.

2.2.1 Total Factor Productivity Accounting

Standard economic theory views production as a transformation of a collection of inputs into outputs. We are interested in how this production structure is changing over time. At an aggregate *National Income and Product Accounts* (NIPA) level we deal with some measure of aggregate output, y, and two measures of aggregate inputs: capital, k, and labor, l. The production structure is represented by the production function, F: $y = F(k, l, t)$. Since the production structure may change, the production function is indexed by time, t. Aggregate total factor productivity changes when the production function shifts over time, i.e., when

[13] Our presentation is instrumental to the discussion of the impact of technological change on labor markets, and hence it is kept to the bare minimum. Jorgenson's (2004) chapter of this Handbook provides an exhaustive treatment of traditional and modern growth accounting.

there is a change in output which we cannot attribute to changes in inputs. More formally, the marginal change in output is the sum of the marginal changes in inputs, weighted by their marginal contributions to output (marginal products), and the shift of the production function, $\dot{y} = F_k\dot{k} + F_l\dot{l} + F_t$.[14] This is usually expressed in terms of growth rates as

$$\hat{y} = \eta_k\hat{k} + \eta_l\hat{l} + \hat{A}, \text{ with } \hat{A} = F_t/F, \tag{2}$$

where hats denote percentage growth rates, and the weight on an input growth rate is the elasticity of output with respect to the input: $\eta_k = F_k k/F$ and $\eta_l = F_l l/F$. Alternatively, if we know the elasticities, we can derive productivity growth as output growth minus a weighted sum of input growth rates.

Solow's (1957) important insight was that, under two assumptions, we can replace an input's output elasticity—which we do not observe—with the input's share in total revenue, for which we have observations. First, we assume that production is constant returns to scale, i.e., that if we are to double all inputs, then output will double, implying that the output elasticities sum to one: $\eta_k + \eta_n = 1$. Second, we assume that producers act competitively in their output and input markets, i.e., that they take the prices of their products and inputs as given. Profit maximization then implies that inputs are employed until the marginal revenue product of an input is equalized with the price of that input. In turn, this means that the output elasticity of an input is equal to the input's revenue share. For example, for the employment of labor, profit maximization implies that $p_y F_l = p_l$, which can be rewritten as $\eta_l = F_l l/F = p_l l/p_y y = \alpha_l$ (p_i stands for the price of good i). With these two assumptions, we can calculate aggregate productivity growth, also known as total factor productivity (TFP) growth, as

$$\hat{A} = \hat{y} - (1 - \alpha_l)\,\hat{k} - \alpha_l\hat{l}. \tag{3}$$

The Solow growth accounting procedure has the advantage that its implementation does not require very stringent assumptions with respect to the production structure, except con-

[14] The marginal change of a variable is its instantaneous rate of change over time; that is, if we write the value of a variable at a point in time as $x(t)$, then the marginal change is the time derivative, $\dot{x}(t) = \partial x(t)/\partial t$. Nothing is lost in the following if the reader interprets $\dot{x}(t)$ as the change of a variable from year to year; that is, $x(t) - x(t-1)$.

stant returns to scale, and it does not require any information beyond measures of aggregate output and input quantities and the real wage. This relatively low information requirement comes at a cost: this aggregate TFP measure does not provide any information on the specific sources or nature of technological change.

Given available data on quantities and prices for industry outputs and inputs, it is straightforward to apply the Solow growth accounting procedure and obtain measures of sector-specific technical change (see, for example, Jorgenson et al., 1987). Recently Jorgenson and Stiroh (2000) have documented the substantial differences in output and TFP growth rates across U.S. industries over the period 1958-1996. In particular, they point out that TFP growth rates in high-tech industries producing equipment investment are about three to four times as high as a measure of aggregate TFP growth. Also based on industry data, Oliner and Sichel (2000) and Jorgenson (2001) attribute a substantial part of the increase of aggregate TFP growth over the second half of the 1990s to one industry: semi-conductors.

2.2.2 Sector-Specific Productivity Accounting

The convincing evidence for persistent differences of SSP growth raises the potential of serious aggregation problems for the analysis of aggregate outcomes. We now discuss SSP accounting in a simple two-sector growth model that focuses on the distinction between investment and consumption goods. This approach provides a straightforward measure of SSP growth, and it keeps the aggregation problems manageable. Based on this approach, we present evidence of substantial increases of the relative productivity in the equipment-investment goods producing industries and stagnant productivity in the consumption goods industries since the mid 1970s.

Greenwood et al. (1997) use a two-sector model of the economy—where one sector produces consumption goods and the other new capital—to measure the relative importance of total-factor productivity changes in each of these sectors. Goods —consumption, c and new capital, x—are produced using capital and labor as inputs to constant-returns-to-scale technologies,

$$c = A_c F(k_c, l_c) \text{ and } x = A_x F(k_x, l_x); \tag{4}$$

and total factor inputs can be freely allocated across sectors,

$$k_c + k_x = k \text{ and } l_c + l_x = l. \tag{5}$$

Note that we have assumed that factor substitution properties are the same in the two sectors; that is, the functions relating inputs to outputs are the same. One can show that with identical factor substitution properties, the two-sector economy is equivalent to a one-sector economy with exogenous changes in the relative price of investment goods, $1/q$

$$y = c + x/q = A_c F(k, l). \tag{6}$$

In particular, the relative price of investment goods is the inverse of the relative productivity advantage of producing new capital goods:[15]

$$q = A_x/A_c. \tag{7}$$

The relative productivity of the investment goods sector is also called "capital-embodied" technical change, because q can be interpreted as the productivity level (quality) embodied in new vintages of capital.[16]

Accounting for quality improvements in new products is a basic problem of growth accounting.[17] This is especially true for our framework since we measure investment in terms of constant-quality capital goods. In a monumental study, Gordon (1990) constructed quality-adjusted price indexes for different types of producers' durable equipment. Building on Gordon's work, Hulten (1992), Greenwood et al. (1997), and Cummins and Violante (2002) have derived aggregate time series for capital-embodied technical change in the U.S. economy.[18] They use the property just described: that the constant-quality price of investment

[15] Jorgenson (2004), in this handbook, labels this methodology, where *relative productivity* growth is measured off the decline in relative prices, the "price approach" to growth accounting.

[16] Define investments in consumption units as $i = x/q$. Then, the aggregate resource constraint reads

$$c + i = A_c F(k, l),$$

and the law of motion for capital in efficiency units is $k' = (1 - \delta) k + iq$.

[17] See this Handbook's chapter by Bils and Klenow (2004) on the measurement of quality for an overview of the different approaches.

[18] Hulten's series strictly uses Gordon's data and therefore spans until 1983. Greenwood et al. extend

relative to consumption (precisely, nondurable consumption and services) reveals the extent of productivity improvements. Their main finding is that:

Observation 12 Productivity growth in the sector producing equipment investment has accelerated relative to the rest of the economy since the early to mid-1970s.

The solid line in Figure 2 shows the relative productivity of the equipment investment goods sector, q, for the period 1947-2000, normalized to 1 in the first year. This index grows at an annual rate of about 1.6 percent until 1975 and at an annual rate of 3.6 percent thereafter. In the 1990s, productivity growth embodied in capital has been spectacularly high, reaching an average annual rate just below 5 percent.

FIGURE 2

The measurement of SSP growth through changes in relative prices requires that the price measures used are appropriately adjusted for quality improvements, presenting a problem for the time period studied since, arguably, the IT revolution has caused large improvements in the quality of durable goods and has led to the introduction of a vast range of new items. Therefore, alternative ways of measuring capital-embodied productivity advancements have been proposed. Hobijn (2000) calculates the rate of embodied technical change by calibrating a vintage capital model. His findings are very similar to the price-based approach, both in terms of the average growth rate, and in terms of the timing of the technological acceleration. Bahk and Gort (1993) and Sakellaris and Wilson (2004) use plant-level data to estimate production functions and assess the productivity effects of new investments. They estimate the growth rate of capital-embodied technical change to be between 12 and 18 percent per year, much higher than the rest of the literature.

Gordon's index to 1992 by applying a constant adjustment factor to the National Income and Product Accounts (NIPA) official price index. Cummins and Violante update the series to 2000. Starting with Gordon's quality-adjusted price indexes for a variety of equipment goods from 1947 to 1983, they estimate the quality bias implicit in the NIPA price indexes for that period. Using the official NIPA series, they then extrapolate the quality bias from 1984 to 2000 for each equipment type and aggregate into an index for equipment and structure.

We calculate the rate of SSP change in the consumption goods sector based on the standard Solow approach. It is well known that the U.S. labor income share in GDP has been remarkably stable for the time period considered. We therefore choose a Cobb-Douglas parametric representation of the production function,

$$y = A_c k^{\alpha} l^{1-\alpha}, \tag{8}$$

with labor income share, $1 - \alpha = 0.64$ (Cooley and Prescott 1995). Conditional on observations for real GDP (in terms of consumption goods), the real capital stock, and employment, we can use this expression to solve for the SSP of the consumption sector A_c.[19] The common finding from this computation, as evident from the dashed line in Figure 2, is

Observation 13 Productivity in the sector producing consumption goods (precisely, non-durable and services) shows essentially no growth over the two decades 1975-1995.

The approach of Greenwood et al. (1997) defines aggregate output in terms of consumption goods. This is rather non-standard. The usual approach, especially as applied to the study of SSP, defines aggregate output growth as a revenue-weighted sum of sectoral output growth rates: a Divisia index (see, e.g., Jorgenson 2001, or Oliner and Sichel 2000). For this more standard approach, one can write aggregate TFP growth as the revenue-weighted sum of sectoral TFP growth. While the Divisia-aggregator approach is a definition with some desirable properties, the Greenwood et al. (1997) approach is based on a particular theory and requires certain identifying restrictions concerning the production structure. Hall (1973) shows that in multi-sector models a unique output aggregator, that is, a function that relates some measure of aggregate output to some measure of aggregate input, exists if

[19] It is important to adjust the capital and labor input measure for quality change. As pointed out above, quality adjustment of investment is useful so as to capture investment-specific technical change. The capital stock is then calculated as the cumulative sum of past undepreciated constant-quality investment. From our discussion of wage inequality it follows that the labor input needs to be adjusted for two reasons. First, the skill premium has been increasing since the mid-1970s, and thus the productivity of skilled labor, A_s, is increasing faster than the productivity of unskilled labor, A_u. Second, at the same time, the relative supply of skilled labor has been increasing, inducing large changes in the composition of the stock of labor. To account for quality changes, we use the labor input index computed by Ho and Jorgenson (1999). The dotted line in Figure 2 plots this quality index for labor which grows at an average rate of 0.8% per year.

certain separability conditions for the aggregate production possibility frontier are satisfied. The conditions for such an output aggregator to exist are, essentially, the ones imposed by Greenwood et al. (1997).[20] Given the definition of aggregate output in equation (6), SSP for consumption, or A_c, is then sometimes interpreted as neutral, or disembodied, aggregate technological change.

2.2.3 Reconciling the Acceleration in Investment-Specific Productivity Growth with the Slowdown in TFP: General Purpose Technology and Learning

The stagnation of aggregate TFP since the mid-1970s—evident from Figure 2—accounts for the phenomenon often referred to as a "productivity slowdown" in the growth accounting literature.[21] How can we reconcile the acceleration of investment SSP with a slowdown of consumption SSP? One interpretation builds on learning-by-doing (LBD). New investment goods do not attain their full potential as soon as they are introduced but, rather, their productivity can stay temporarily below the productivity of older capital that was introduced same time ago. This feature is attributed to learning effects.[22]

These learning effects can be extremely important when the technological change is "drastic." Recent discussions suggest that the advent of microelectronics led to a radical shift in the technological paradigm, i.e., to a new "general purpose technology" (GPT). Bresnahan and Trajtenberg (1995) coined this term to describe certain major innovations that have the potential for pervasive use and application in a wide range of sectors of the economy. David (1990) and Lipsey et al. (1998) cite the microchip as the last example of such innovations that included, in ancient times, writing and printing and, in more recent times, the steam-engine and the electric dynamo.[23] Although it is hard to define the concept satisfactorily,

[20]For further details on this issue, see Hornstein and Krusell (2000).

[21]Since non-equipment investment represents more than three-fourths of GDP, the slowdown of consumption SSP change accounts for most of the slowdown of aggregate TFP change.

[22]The literature on learning effects is large. Lucas (1993) discusses the classic example of LBD related to the construction of the liberty ships of World War II. Bahk and Gort (1993) measure substantial LBD effects at the plant level. Irwin and Klenow (1994) present evidence of LBD in the production of semiconductors. Jovanovic and Nyarko (1995) document learning curves in several professions. Huggett and Ospina (2001) find evidence that the effect of a large equipment purchase is initially to reduce plant-level total factor productivity growth.

[23]Gordon (2000) offers a dissenting view on the issue of whether or not information technologies measure up to the great inventions of the past. In his view, the aggregate productivity impact of computers

given available data, we list as a "fact" the dominant view, which maintains that:

Observation 14 Technological change in the past 30 years displays a "general purpose" nature.

Though most of the evidence supporting this statement is anecdotal, there are some bits of hard evidence. Hornstein and Krusell (1996) document that the decline in TFP occurred roughly simultaneously across many developed countries. More recently, Cummins and Violante (2002) construct measures of productivity improvements for 26 different types of equipment goods. Using the sectoral input-output tables, they aggregate these indexes into 62 industry-level measures of equipment-embodied technical change, and document that their growth rate has accelerated by a similar amount in virtually every industry in the 1990s. Jovanovic and Rousseau (2004a) draw an articulated parallel between the diffusion of electricity in the early 20th century and the diffusion of information technologies (IT) eighty years later based on a variety of data. Their evidence supports the view that both episodes marked a drastic discontinuity in the historical process of technological change. Taken together, all these observations suggest that, similar to other past GPTs, IT has affected productivity in a general way over the past three decades.

There are two versions of the argument that IT are responsible for the observed productivity slowdown. According to one, the slowdown is real: when learning-by-doing is important in improving the efficiency of a production technique, abandoning the older, but extensively used technology to embrace a new method of production involves a "step in the dark" that can lead to a temporary slowdown in labor productivity (Hornstein and Krusell, 1996, Greenwood and Yorukoglu, 1997, and Aghion and Howitt, 1998, chapter 8).

An alternative, complementary version maintains that the slowdown is a statistical artifact due to mismeasurement: if the phase of IT adoption coincides with associated investments in organizational or intangible capital, as our Section 5 will suggest, then insofar as these investments are not included in the official statistics, measured TFP growth will first

and telecommunications equipment has been fairly small compared to, say, the telegraph, the railroad, or electricity.

underestimate and then overestimate "true" TFP growth (Hall, 2001, and Basu et al. 2003). The reason is that initially, when large investments in organizational capital are made, the "output side" of the mismeasurement is severe. Later, when the economy has built a significant stock of organizational capital, the "input side" of mismeasurement becomes dominant.

This explanation of the TFP slowdown is appealing, but extremely difficult to evaluate quantitatively because of the lack of direct evidence on how organizations learn. Using some theory, Hornstein (1999) argues that one key parameter is the fraction of knowledge that firms can transfer from the old to the new technology but also shows that the model's predictions vary significantly across plausible parameterizations. Atkeson and Kehoe (2002) build an equilibrium model to measure the dynamics of organizational capital during the "electrification of America". They criticize the Bahk and Gort (1993) view that organizational learning is reflected into an increase in the productivity of labor at the plant level: in an equilibrium model where labor is mobile, productivity is equalized across plants. Instead, they argue that when organizations learn they expand in size. Thus, cross-sectional microdata on the size distribution of plants allows to identify the structural parameters of the stochastic process behind organizational learning.

Finally, Manuelli (2000) argues that, even in absence of learning effects, the anticipation of a future technological shock embodied in capital can result in a transitional phase of slowdown of economic activity. In the period between the announcement and the actual availability of the new technology, the existing firms prefer exercising the option of waiting to invest and the new firms prefer delay entering. Consequently, output falls temporarily until the arrival of the new technology.[24]

2.2.4 Factor-Specific Productivity Accounting

In order to talk about changes in FSP, one possibility is to generalize the production function in equation (8) by disaggregating the contributions to production of the two labor inputs—skilled (e.g., more educated) and unskilled (e.g., less educated) labor. Suppose the aggregate

[24]We refer the reader to Hornstein and Krusell (1996) for a list of alternative explanations of the TFP slowdown that are not based on changes in technology.

labor input, l, is a CES function of skilled and unskilled labor, l_s and l_u, with FSPs A_s and A_u:

$$l = \left[(A_s l_s)^{\sigma} + (A_u l_u)^{\sigma} \right]^{1/\sigma}, \ \sigma \leq 1. \tag{9}$$

Relative wage data can then be employed to understand the nature and evolution of FSP in the economy. With competitive input markets, the relative wages are a function of the relative FSP and the relative labor supply:

$$\ln \left(\frac{w_s}{w_u} \right) = \sigma \ln \left(\frac{A_s}{A_u} \right) - (1 - \sigma) \ln \left(\frac{l_s}{l_u} \right). \tag{10}$$

The elasticity of substitution between skilled and unskilled labor here is $1/(1-\sigma)$. Katz and Murphy (1992) run a simple regression of relative wages on relative input quantities and a time trend to capture growth in the ratio $\frac{A_s}{A_u}$. They measure skilled labor input as total hours supplied to the market by workers with at least a college degree. Their estimate of the substitution elasticity—around 1.4 (or $\sigma = 0.29$)—indicates that a ten-percent increase in the relative supply of skilled labor implies a seven percent decline of the skill premium.[25] The estimated elasticity of substitution between factors, together with the growing skill premium, imply an increase in the relative FSP of skilled labor in excess of 11 percent per year. We conclude that the typical result of similar exercises on U.S. data is that:

Observation 15 Recent technological advancements have been favorable to the most skilled workers in the population. In other words, technical change has been *skill-biased*.

The "acceleration" in the rate of capital-embodied technical change, the "general purpose" nature of the new wave of technologies, and the "skill-biased" attribute of the recent productivity advancements are the three chief features of the new technological environment that seems to have emerged since the early to mid 1970s. The various economic theories that we are about to review in the rest of this chapter are built on various combinations of these features.

[25]The estimated input elasticity of about 1.4 is consistent with a large empirical literature on factor substitution that uses a wide array of data sets (time series as well as cross-section) and methods; see, e.g., Hamermesh (1993).

3 Skill-Biased Technical Change: Inside the Black Box

As we have just observed, the pattern of relative quantities of skills measured by education suggest that the behavior of the skill premium, that is, the increase in the wages of highly educated workers relative to those of less educated workers, should be attributed to a skill-biased labor demand shift, or to "skill-biased technical change." In the absence of a factor-bias in technological progress, the upward trend in the supply of skills documented in Figure 1 (top panel) would have reduced the skill premium.

Katz and Murphy (1992) were the first to use a production framework with limited substitution between skilled and unskilled labor to recover changes in relative FSP from changes in the skill premium. One should note a substantial drawback of the pure skill-biased technical change hypothesis: it is based on *unobservables* (relative FSP changes) that are measured residually from equation (10), so very much like TFP, it is a "black box". In this section we review the attempts to give some specific economic content to the notion of skill-biased technical change.

We start from the capital-skill complementarity conjecture advanced originally by Krusell et al. (2000). Next, we analyze models based on the Nelson-Phelps hypothesis: the adoption phase of a new technology requires skilled and educated workers. If one allows for an important role of FSP changes, then it is paramount to understand what economic forces induce these changes endogenously. In this context, we review the theory of "directed technical change" associated mainly to the work by Acemoglu (2002b, 2003b): exogenous spurts in the relative supply of skilled labor can induce the introduction of skill-biased technological advancements by affecting the incentives of the innovators.

3.1 Capital-Skill Complementarity

Krusell et al. (2000; KORV henceforth) argue that the dynamics of SSP that induced the substantial drop in the relative price of equipment capital is the force behind the rise in the skill premium. The decline in the price of equipment due to productivity improvements, especially that embodied in information and communication technologies, led to an increased

use of equipment capital in production. KORV observe that, at least since Griliches (1969), various empirical papers support the idea that skilled labor is relatively more complementary to equipment capital than is unskilled labor. As a result, the higher capital stock increased skilled wages relatively more than unskilled wages. Consequently, the skill premium increased.

Thus, the key elements in KORV's analysis are: 1) separating the effect of equipment capital from that of other capital, mainly structures, 2) allowing equipment to have different degrees of complementarity with skilled and unskilled labor, 3) measuring the efficiency units of capital, especially the new technologies, correctly.[26]

KORV capture the differential complementarity between capital and skilled and unskilled labor using the following nested CES production function of four inputs: structures, k_s; equipment, k_e; skilled labor, l_s; and unskilled labor, l_u:

$$y = k_s^{\alpha}\left[\lambda\left[\mu\left(A_{k_e}k_e\right)^{\rho} + (1-\mu)\left(A_s l_s\right)^{\rho}\right]^{\sigma/\rho} + (1-\lambda)\left(A_u l_u\right)^{\sigma}\right]^{\frac{1-\alpha}{\sigma}}, \tag{11}$$

with $\rho, \sigma \leq 1$. Profit-maximizing behavior of a price-taking firm implies that the skill premium can be approximately written as

$$\ln\left(\frac{w_s}{w_u}\right) \simeq \sigma \ln\left(\frac{A_s}{A_u}\right) - (1-\sigma)\ln\left(\frac{l_s}{l_u}\right) + \lambda\frac{\sigma-\rho}{\rho}\ln\left(\frac{k_e}{l_s}\right)^{\rho}. \tag{12}$$

KORV estimate $\sigma = 0.4$ and $\rho = -0.5$, and thus that the skill premium increases with the stock of equipment capital.[27] They find that the relative productivity of skilled labor grows

[26]The quality-adjusted equipment capital stock is again based on the work of Gordon (1990) and subsequent updates, especially for IT technology.

[27]With this nested CES in 3 factors (equipment, skilled and unskilled labor) it is unclear how to define capital-skill complementarity. One possible, but slightly unhortodox, definition is that the skill premium rises with the stock of equipment. A more traditional definition involves comparing the Allen elasticities of substitution. The elasticity of substitution between equipment and unskilled labor is $1/(1-\sigma)$, while the elasticity of substitution between equipment and skilled labor is

$$\frac{1}{1-\sigma} + \left(\omega_e + \omega_s\right)^{-1}\left[\frac{1}{1-\rho} - \frac{1}{1-\sigma}\right],$$

where ω_e and ω_s are, respectively, the income shares of equipment and skilled labor. Thus, according to both definitions, the parameter estimates in KORV imply that equipment capital is more complementary with skilled labor compared to unskilled labor. See Ruiz-Arranz (2002) for a discussion of the various definitions of elasticity of substitution in production function with more than 2 inputs. Interestingly, Ruiz Arranz divides equipment into finer categories and finds that IT capital (defined as computers, communication equipment and software) is the subgroup with the largest degree of capital-skill complementarity.

at a modest 3 percent per year, a much more plausible number than the one estimated by Katz and Murphy (1992). Overall, KORV show that with their estimated parameters, the relative wage movements in the data can be quite closely tracked. This includes the decline of the wage premium in the 1970s, attributable to an acceleration in the growth of college enrollment due to the Vietnam war draft and the entry of the baby-boom cohorts.[28]

From equation (12) it follows that the skill premium can increase, even if the relative productivity of skilled labor remains constant and the relative supply of skilled labor increases, provided the equipment-skilled labor ratio trends upward fast enough. From this perspective, the results of KORV complement Katz and Murphy's (1992) work: when capital and skills are complementary in production, capital accumulation can explain a large fraction of the residual trend in skill-biased productivity growth.[29]

3.1.1 Further Applications of the Capital-Skill Complementarity Hypothesis

In KORV, the production structure is "centralized" through an aggregate production function. Jovanovic (1998) models an economy with vintage capital where production is decentralized into machine-worker pairs. Newer machines are more productive than older machines, and workers differ in their innate skill level. The pair's output is a multiplicative function of these two inputs. Jovanovic assumes perfect information (no coordination frictions), and hence the labor market equilibrium assignment displays "positive sorting" between skills and machines' productivity (Becker, 1973), i.e., capital-skill complementarity emerges endogenously.[30] An acceleration in the growth rate of technology embodied in

[28]Lee and Wolpin (2004) find evidence of capital-skill complementarity both in the goods-producing industries and services in the U.S. economy, and argue that it is an important ingredient to explain the pattern of relative wages and relative labor inputs across the two sectors, over the past 50 years.

[29]Acemoglu (2002a) argues that if the capital-skill complementarity hypothesis is valid, then in equation (10) the relative price of equipment should proxy the shift in the demand for skills and perform better than a linear time trend. However, he finds the trend is always more significant. First, as equation (12) shows, the right variable to add to the Katz-Murphy equation is not the relative price of equipment, but the equipment-skill labor ratio. Second, even using this latter variable one would be bound to find that the linear time trend is more significant because in an OLS regression the estimated coefficient on the time trend converges to its true value at a faster rate than the coefficient on the equipment-skill ratio. More importantly, the key insight of KORV is to give an economic content to the "skill-biased technical change" view, by replacing an unobservable trend with an observable variable.

[30]Holmes and Mitchell (2004) start from a more primitive level where production combines tasks of various complexity and the production factors can perform tasks at a given setup-cost per task. They show that

machines, that is, an increase of the relative productivity differences across vintages, has two effects: 1) for a given age range of machines, it widens the underlying distribution of job productivity differences and, since in equilibrium high-skilled workers are assigned to high-productivity machines, it magnifies the skill premium; 2) the faster rate of obsolescence shortens the optimal life of capital, that is, the range of operative vintages narrows, which tends to weaken inequality since the least productive workers are now matched with better machines. As we will see, these two counteracting forces will survive in the frictional economies of section 7, in spite of the different nature of the equilibrium assignment of workers to machines.

The capital-skill complementarity hypothesis has proved to be helpful to interpret the dynamics of the skill premium in other countries. Perhaps the most interesting example is Sweden. Lindquist (2002) documents that the facts to be explained in Sweden are qualitatively similar to the U.S. facts: between 1983 and 1999 the college premium rose by over 20% and the supply of skilled workers increased substantially. Sweden represents an especially interesting test case for the KORV model because Swedish labor market institutions are commonly believed to play a crucial role in wage setting, arguably making market forces less critical in determining relative wage movements. The main result of Lindquist (2002) is that capital-skill complementarity explains close to half of the dynamics of the skill premium.[31]

How can one reconcile the traditional strength of labor market institutions, such as unions and collective bargaining, in the Swedish labor market with the finding that market forces account for a large part of relative wage dynamics? One possibility is that institutions set the aggregate share of income going to labor in any given period—possibly extracting rents from firms. The distribution of these rents among workers is then determined by their individual outside options, which differ across skill levels and are affected by technical change. In section 8 we develop further this conjecture in the context of the decline in union membership in the United States, but the economic linkages between the dynamics of

under reasonable primitive assumptions on setup costs for capital, skilled labor and unskilled labor, the former two inputs display a form of complementarity.

[31] Lindquist uses the KORV specification for aggregate technology in equation (11) and estimates $\rho = -0.92$ and $\sigma = 0.31$.

institutions and technological progress are far from being well understood.

More international evidence in favor of the capital-skill complementarity model is offered by Flug and Hercowitz (2000). They estimate a strong effect of equipment investment on relative wages and employment of skilled labor using a panel data set for a wide range of countries around the world.

Recently, the capital-skill complementarity idea has been imported into the study of inequality at the business-cycle frequency. The skill premium is found to be close to *acyclical* in the United States: its contemporaneous correlation with output is positive, but not statistically different from zero. Lindquist (2004) argues that, since unskilled labor is relatively more pro-cyclical than skilled labor, a Cobb-Douglas production function in three inputs (capital, skilled labor, and unskilled labor) would predict a strongly pro-cyclical skill premium. Inspection of equation (12) suggests that introducing capital-skill complementarity in production can help matching the data since, at impact, skilled hours respond more than the stock of equipment: the capital-skill complementarity effect is countercyclical and offsets the change in relative supply.[32]

In sum, the studies discussed in this section indicate that capital-skill complementarity is a quantitatively important ingredient in competitive theories of relative wage determination, within centralized as well as decentralized production structures and at high as well as low frequencies.

3.2 Innate Skills and the Nelson-Phelps Hypothesis

Nelson and Phelps (1966) argued that the wage premium for more skilled workers is not just the result of their having higher "static productivity". Workers endowed with more skills, they contended, tend to deal better with technological change in the sense that their productivity is less adversely affected by the turmoil created by technological transformations of the workplace, or in that it is less costly for them to acquire the additional skills needed to use a new technology. Greenwood and Yorukoglu (1997) cite sources reporting that the skill

[32]Within a similar framework, Cohen-Pirani and Castro (2004) argue that capital-skill complementarity is important for understanding why the volatility of skilled labor (relative to GDP) has tripled after 1985.

premium also rose during the course of the first industrial revolution. In the context of the recent "IT revolution", Bartel and Lichtenberg (1987) provide evidence that more educated individuals have a comparative advantage at implementing the new technologies and and Bartel and Sicherman (1998) argue that high-skilled workers sort themselves into industries with higher rates of technical change.

The theory has been formalized in various formats. Lloyd-Ellis (1999) embeds a race between the innovation rate and the "technological absorption rate" of workers (the maximum numbers of innovations that can be adopted per unit of time) in a general equilibrium model: at times when the innovation rate exceeds the absorption rate, wage inequality increases due to the fierce competition for scarce, adaptable labor. Galor and Moav (2000) formalize this hypothesis differently and assume that technological change depreciates the human capital of the unskilled workers faster than that of skilled workers (the "erosion effect"). Krueger and Kumar (2004) distinguish between workers with general education and workers with vocational skill-specific education and postulate that only the former type remains productive when new technologies are incorporated into production.

It is important to remark that this hypothesis, in all its versions, applies to educational skills as well as dimensions of skills that are not necessarily observable or correlated with education. Hence, it can potentially account for the rise in within-group (or residual) inequality. Ingram and Neumann (1999) offer some evidence on the increase in the return to certain categories of skills not fully captured by education. They match individual data on wages and occupations from the CPS with the skill content of several occupations, obtained from the *Dictionary of Occupational Titles* (DOT). DOT data contain information on how much each occupation requires of each of a wide range of skills such as verbal aptitude, reasoning development, numerical ability, motor coordination, and so on. Using factor analysis they group over 50 type of skills into four factors (intelligence, clerical skills, motor skills, and physical strength) and estimate that the return to "intelligence" has almost doubled from 1971 to 1998. Moreover, adding the quantity of this factor to a standard Mincerian wage regression weakens the implied increase in the returns to college education significantly.

The idea that the diffusion of IT may have raised the demand for adaptable skilled workers—thus, even within educational groups—has been formalized in various ways by Galor and Tsiddon (1997), Greenwood and Yorukoglu (1997), Caselli (1999), Galor and Moav (2000), and Aghion et al. (2002).

To illustrate the basic mechanism of such a theory, consider an economy where workers differ in their cost of learning the new technology.[33] Suppose that this economy starts in a steady-state equilibrium where production uses the "old" technology, $y_1 = A_1 k_1^{\alpha} l_1^{1-\alpha}$. The labor market is competitive; thus, in steady state, all workers are employed in the old sector and there is no wage inequality.

Suppose a new technology becomes available and the sector using this new technology can produce output with $y_0 = A_0 k_0^{\alpha} l_0^{1-\alpha}$ where $A_0 > A_1$. Because of the learning cost, labor is not perfectly mobile, and wages in the two sectors may differ. Capital, however, is free to move towards its more productive use, and factor-price equalization for capital yields

$$R_0 = R_1 \Rightarrow \frac{l_1}{k_1} = \left(\frac{A_0}{A_1}\right)^{\frac{1}{1-\alpha}} \frac{l_0}{k_0}. \tag{13}$$

It is straightforward to show that

$$w_0 = \left(\frac{A_0}{A_1}\right)^{\frac{1}{1-\alpha}} w_1 > w_1.$$

Therefore, in equilibrium, a premium emerges for those workers with low learning cost (i.e., high ability) who can adapt quickly and move to the new sector.

The skill premium increases due to two effects. With full mobility of labor, inequality would disappear. With no labor mobility and no capital mobility, the skill premium would reflect the productivity difference A_0/A_1. In this class of models, labor mobility is limited by the distribution of ability in the economy, but capital moves freely. Full mobility of capital induces a general equilibrium feedback that amplifies inequality: factor-price equalization requires capital to flow to the sector operating the new technology to equate marginal productivities of capital.[34] Thus, workers on the new technologies are endowed with more capital, which boosts their relative wages further.

[33] In the rest of this section, the exposition will be based mainly on the environment in Caselli (1999).

[34] One implication of this mechanism, evident from equation (13), is that a technological revolution should

In its typical version, the Nelson-Phelps hypothesis implies that the rise in the skill premium will be transitory: it is only in the early adoption phase of a new technology that those who adapt more quickly can reap some benefits. Over time there will be enough workers who learn how to work with the new technology to offset the wage differential. Note the difference with the KORV hypothesis, where the effect on the skill premium is permanent. Are new technologies and skills complement in the whole production process or just in the adoption phase? Chun (2003) uses industry-level data for the U.S. to disentangle the impact of "adoption" and "use" of IT. He finds that the increase in the relative demand of educated workers from 1970 to 1996 in the U.S. is related significantly to both factors, but quantitatively the impact of use is twice as large.

3.2.1 Further Applications of the Nelson-Phelps Hypothesis

Aghion (2002) and Borghans and ter Weel (2003) emphasize that the Nelson-Phelps approach can explain why, in the 1970s, the college premium declined at the same time that the wage dispersion within college graduates increased. The idea is that in the early phase of IT diffusion in the 1970s only educated workers with high ability adopt. Naturally, this higher return to ability increases within-group inequality. The contemporaneous acceleration in the growth of the supply of educated labor, due to exogenous factors, explains the relative fall in the average wage of college graduates.

Beaudry and Green (2003) compare the United States and Germany, highlighting an apparently puzzling feature of the data: the relative supply of skilled labor in the United States grew faster than in Germany, and yet the skill premium rose in the United States, but not in Germany. They outline a model that combines elements of Caselli (1999) and Krusell et al. (2000). Consider an economy where there are two technologies in operation and the "new" technology displays more capital-skill complementarity than the old one. An exogenous rise in the supply of skills increases the relative return to capital in the new sector. Capital then flows from the old to the new sector, and, ultimately, this higher capital

trigger a surge in the real rate of return on capital by a factor $(A_0/A_1)^{\frac{1}{1-\alpha}}$. Yearly U.S. long-term real interest rates were roughly 3 percent higher in the period 1980-1995 compared to the period 1965-1980. It is unclear whether this magnitude is quantitatively consistent with the observed increase in wage dispersion.

intensity can raise the relative wage of skilled labor if labor is not perfectly mobile because, as in Caselli's model, only skilled workers can quickly adapt. Thus, in the long-run, the country with the initial spur in the supply of skilled labor (the United States) finds itself with a larger skill premium.

In their original paper, Nelson and Phelps (1966) developed the concept of "technological gap", defined as the percentage difference between the technology operated by the typical machine in the economy and the one embodied in the leading-edge machine. They conjectured that a rise in the technological gap should be associated with a large skill premium because of the surge in the demand for educated workers needed to adopt the new, more productive technologies. Cummins and Violante (2002) use data on the quality-adjusted relative prices and quantities of equipment investment to construct a measure of the technological gap for the U.S. economy.[35] Figure 3 shows that the technological gap and the skill premium have moved largely in tune over the past half century, confirming—at least in the time-series dimension—even the most literal version of the Nelson-Phelps hypothesis.

FIGURE 3

Put differently, the size of the technological gap can be thought of a proxy for shifts in the relative demand of skilled workers.

3.3 Endogenous Skill-Biased Technical Change

In the literature we discussed so far, the sector bias and the factor bias of technical change were assumed to be exogenous. Over the past 20 years a substantial body of work in the field of growth theory has formalized the idea that the efforts of innovators are endogenous and

[35] Precisely, if q_t is the level of productivity embodied in the new investment at time t, then the average unit of productive capital in the economy at time t embodies a technology with productivity Q_t, defined as

$$Q_t = \sum_{j=0}^{\infty} (1-\delta)^j \, q_{t-j} \frac{i_{t-j}}{k_t},$$

where δ is the depreciation rate, i denotes investments and k the capital stock, both expressed in units of consumption. In other words, Q_t is the ratio between capital stock correctly measured in efficiency units (the numerator) and capital stock k not adjusted for quality. Then, the gap is defined as $\frac{q_t - Q_t}{Q_t}$.

respond to market incentives. The models belonging to the so-called "new growth theory" describe the endogenous determination of the *level* of innovative activity.

Recently, Acemoglu (1998, 2002b, 2003b) and Kiley (1999) have developed the idea that the composition, or *direction*, of innovations is also endogenous: if R&D activity can be purposefully directed towards productivity improvements of different inputs (capital, skilled labor, and unskilled labor), then it will be biased towards the factor that ensures the largest returns.

An important ingredient of this approach is that the returns to R&D targeted toward a given input are proportional to the total supply of that input, since "productivity" and "quantity" are complements in production. This creates a "market size" effect of R&D: productivity-improving resources are allocated to factor markets with large relative factor supplies.[36]

It is useful to see how this mechanism works within a simple model that represents a reduced form of the richer environments offered by Acemoglu and Kiley. Consider an economy with a given endowment of skilled and unskilled labor, l_s and l_u, and a production function (9) as in Section 2.2.4. Conditional on the FSPs, A_s and A_u, wages and employment are determined competitively, and the competitive equilibrium is Pareto-optimal. Now suppose that the Social Planner wants to maximize production subject to a given frontier of technological possibilities, that is, choices of A_s and A_u:

$$\max_{\{A_s, A_u\}} \left[(A_s l_s)^\sigma + (A_u l_u)^\sigma\right]^{1/\sigma}$$

$$s.t. \qquad \left[\lambda A_s^\phi + (1-\lambda) A_u^\phi\right]^{1/\phi} = 1.$$

Assume that the technological frontier is convex, that is, that $\phi > 1$. Rearranging the first-order conditions, one arrives at

$$\frac{A_s}{A_u} = \frac{\lambda}{1-\lambda}\left(\frac{l_s}{l_u}\right)^{\sigma/(\phi-\sigma)}, \tag{14}$$

which describes the optimal choice of skill-bias given the relative factor supply. The above equation shows that when skilled and unskilled labor are substitutes, $0 < \sigma \leq 1$, the skill

[36] It is useful here draw a parallel with certain traditional endogenous growth models, where the scale effect determines the level of the growth rate. See, Jones (2004) for a survey of the models with scale effects.

bias is increasing in the relative supply of skills. This latter parametric condition implies that the marginal product of each innovation is increasing in its corresponding factor.

A surge in the relative endowment of skilled labor, like the one witnessed by the U.S. economy in the postwar period, induces the adoption of more skill-biased technologies in production. This force tends to counteract the direct relative supply effect on wage inequality. Can the endogenous skill bias be so strong in the long run as to overturn the initial supply effect?

To answer this question, we substitute the expression for the skill bias, (14), into the expression for the skill premium, (10), and obtain

$$\ln\left(\frac{w_s}{w_u}\right) \propto \frac{\sigma - \phi\,(1-\sigma)}{\phi - \sigma} \ln\left(\frac{l_s}{l_u}\right).$$

We see that the skill premium is increasing in the relative supply of skilled labor as long as $\phi \in \left(\sigma, \frac{\sigma}{1-\sigma}\right)$. Thus, theoretically, it is possible to explain a positive long-run relationship between the relative supply of skilled labor and the skill premium as the one depicted in Figure 1 (top panel).

One limitation of existing models of directed technical change, and also of most of the literature surveyed in this section, is that arguments for the skill-premium focus on the response of a relative price to exogenous changes in relative factor supplies. Whereas one can reasonably assume that "ability" is largely pre-determined with respect to the point in the life-cycle when agents start making economic decisions, education is not. One would expect that changes in returns to education as large as the ones we observed in the past 30 years would significantly affect the incentives to acquire education. However, it is an open question to what extent the observed changes in returns were predicted by the cohorts affected by these returns when they made their education decisions.[37]

Models of directed technical change, augmented by an endogenous supply of skills, can give rise to multiple steady states. If the innovators expect the supply of educated workers to rise, they will invest in skill-biased R&D which, in turn, will augment the returns to college and induce households to acquire human capital, fulfilling the innovators' expectations.

[37] See Abraham (2003) for a related analysis.

3.3.1 Sources of the Skill-Bias in Recent Times

Equation (14) shows that the most natural candidate as engine of the recent skill-biased technical change was the rise in the relative supply of educated workers. The latter was, according to Acemoglu (2002a) largely exogenous, at least initially, and a result of the high college enrollment rates of the baby-boom cohort and of the Vietnam war draft. The crucial issue, still unresolved, is whether the necessary parametric restrictions discussed earlier are plausible, and whether the initial shock is large enough.[38] What other changes in the economic environment can be listed as potential sources of skill-biased innovations?

First, there are possible interactions between capital-skill complementarity and the direction of technical change. Hornstein and Krusell (2003) have taken a first step at incorporating the idea of factor-biased innovations into the KORV explanation of the skill premium. Intuitively, an acceleration in capital accumulation due, for example, to an exogenous fall in the price of capital increases the returns to skill-biased innovations if capital is more complementary with skilled labor. Hence, capital-embodied productivity improvements can be the source of factor-biased technical progress. For a calibrated version of their model, Hornstein and Krusell find that a persistent decline of the relative price of capital results in a temporary, but very persistent, increase of the skill premium. In their model the skill premium not only increases because of capital accumulation (as in KORV), but it also increases because of the endogenously induced spur of skill-biased technical change.

Second, the increased openness to trade can play a role. Using a Schumpeterian growth model, Dinopolous and Segerstrom (1999) argue that if trade liberalization boosts the profitability for monopolistic suppliers by increasing the size of their markets, then resources shift from manufacturing to R&D activities. If, in turn, R&D is a skill-intensive sector, the skill premium rises. This model determines endogenously the level of R&D, but does not display endogenous factor bias in the equilibrium innovation rate. In Acemoglu (2003c), the direction of technical change is related to international trade. A natural assumption about

[38] In the richer model developed by Acemoglu (2002b), this parametric restriction requires that the elasticity of substitution between skilled and unskilled labor be larger than 2. Most of the empirical literature on factor substitutability, however, points at values around 1.5 (Hamermesh 1993).

factor endowments is that in the United States the ratio of skilled to unskilled labor is higher than in the rest of the world. After the U.S. economy opens to trade, the world prices are determined by the aggregate relative factor endowment, and thus skill-intensive goods become relatively more expensive. In the class of models with an endogenous factor bias, factors which produce goods with the highest relative price—and the highest expected profits—will be the target of the largest amount of innovative activity (the "relative price" effect). Thus trade opening induces skill-biased technical change. This mechanism can, under some conditions, explain also the increase in the skill premium in less-developed countries documented, for example, by Robbins (1996).

Third, Cozzi and Impullitti (2004) argue that government policy may also have contributed to the bias in technical change. In the 1980s, U.S. technology policy rapidly shifted its priority from security and defense to economic competitiveness in order to counteract the emerging dominance of Japan in the sectors producing high-tech goods.[39] Within a Schumpeterian growth model, they show that when the government reallocates its expenditures towards the (high-tech) manufacturing goods with the highest potential quality improvement, it creates a market-size effect that can lead to a rise in the innovation rate in those sectors and a net increase in the demand for skilled R&D workers and their wages.[40]

Although we have learned from the above analyses about possible channels influencing the skill premium, there is little work that allows us to quantify each of the channels. A careful calibration and evaluation of a model which incorporates these various channels would be an important first step in this direction.

[39]Japan's share of the high-tech goods markets rose from 7 percent to 16 percent during the period 1970-1990, while at the same time the U.S. share declined from 30 percent to 21 percent. In 1963 government spending on defense represented 1.37 percent of GDP. In 1980, it was down to 0.57 percent.

[40]Like the Dinopolous and Segerstrom (1999) model, strictly speaking, this is not a model of directed-technical change, since skilled labor works only in the R&D sector, and each manufacturing sector employs unskilled labor. However, in a version of the model with endogenous factor-bias and a structure of manufacturing where high-tech goods are produced by skilled labor and low-tech goods by unskilled labor, the shift in technology policy would have the same qualitative effect.

3.4 A Historical Perspective on the Skill Premium

In Section 2 we have observed that, over the last 100 years, wage inequality first declined and then increased, with the turning point somewhere around 1950. Can the theoretical models developed to interpret the increasing wage inequality for the second half of the 20th century also account for the declining wage inequality of the first half of the 20th century?

3.4.1 Capital-Skill Complementarity

Figure 4 plots the relative price of equipment together with the returns to one year of education (both college and high school) since 1929.[41] The pattern is rather striking and is broadly consistent with an explanation based on the capital-skill complementarity hypothesis. During the first half of the century, the price of capital increased which slowed the demand for educated labor and the skill premium. Then around mid-century it started to decline, fostering a strong demand shift in favor of educated labor.

FIGURE 4

This extension of the KORV analysis to the whole 20th century is yet to be performed formally.[42] Thus, before one fully subscribes to this explanation, it is worth discussing the key assumption behind the model. Is it an accurate historical assessment that the introduction of new capital goods has systematically increased the productivity of skilled labor relative to the productivity of unskilled labor? In other words, when can one date the birth of work organizations displaying capital-skill complementarity?

According to Goldin and Katz (1998), until the early 20th century there was no trace of skill-biased technical change; rather, the opposite bias was at work. The origins of capital-

[41] The relative price is computed from series available on the BEA website. In particular, compared to the series discussed previously in the chapter, there are no quality adjustments. As a result, the acceleration which occurred since the mid 1970s is less evident here. The series on the return to education for 1939, 1949, 1959, 1969, 1979, 1989, and 1995 are taken from Table 7 in Goldin and Katz (1999) and interpolated linearly for the missing years in between. The first datapoint for 1929 is obtained by linear interpolation from 1914.

[42] Admittedly, the evidence in Figure 4 is rather indirect. Looking directly at the stock of equipment (unadjusted for quality improvements), its average annual growth rate in the periods 1930-1950, 1950-1980, 1980-2000 is, respectively, 2.2%, 5.0% and 4.2%. However, when quality-adjusted, the growth rate of equipment from 1980-2000 is close to 8% (Cummins and Violante, 2002). See also Hornstein (2004) for a discussion of historical trends of U.S. capital-output ratios.

skill complementarity are associated with the introduction of electric motors, and a shift away from assembly lines and toward continuous and batch processes. This development started in the second and third decades of the 20th century. In particular, the declining relative price of electricity, and the consequent electrification of factories, made it possible to run equipment at a higher speed. This, in turn, increased the demand for skilled workers for maintenance purposes. Since then, the introduction of new equipment, such as numerically controlled machines, robotized assembly lines, and finally computers further increased the relative productivity of skilled labor. Thus, we conclude that based on anecdotal evidence, the period portrayed in Figure 4 is one where capital-skill complementarity became more important.

Mitchell (2001), in a related interpretation on the last century of data, emphasizes the technological aspects of optimal plant size. Mitchell documents a striking similarity between the historical path of wage inequality and the pattern of average plant size in manufacturing which rose over the 1900-1950 period and shrunk between 1950 and 2000, thus almost producing the mirror image of inequality at low frequencies. The time-path of plant size can be interpreted as an indicator of the magnitude of the fixed costs of capital and fits well with the evidence of Figure 4.

In Mitchell's model, production requires performing a large set of tasks with capital and two types of labor, skilled and unskilled. Entrepreneurs face a fixed cost to operate capital, skilled labor, and unskilled labor. Unskilled labor has a higher fixed cost and a lower variable cost than does skilled labor; e.g., unskilled labor is specialized and needs a certain amount of training to perform all the tasks, whereas skilled labor is naturally able in multi-tasking.[43]

The move from craft shops to assembly lines (1900-1950) induced a rise in the fixed cost: the optimal size of the plant rose and with a larger size, plants optimally employed more unskilled workers with large fixed cost, but low variable cost (wages). The demand for

[43]This idea is further developed in Holmes and Mitchell (2004). This paper develops a theory of the intrinsic difference between three key factors of production: capital, unskilled labor, and skilled labor. Based on this theory, the authors develop implications for: 1) how capital and skill intensity vary as a function of size of plants, 2) the micro-foundations of capital-skill complementarity, 3) the effect of trade on the skill premium and the historical relationship between the plant size-skill correlation and the skill-premium.

unskilled workers rose, weakening the skill premium. As an illustration of the importance of fixed costs for this type of production method, recall that all Ford plants had to be closed and redesigned when the "Model T" was discontinued (Milgrom and Roberts, 1990).

The shift toward more flexible, numerically controlled machines and IT capital (1950-2000) led firms to adopt a smaller scale of production and employ more highly skilled workers whose low fixed cost makes them preferable to unskilled workers in small plants. The increased demand for skilled labor thus raised the skill premium. Based on a calibration exercise, the model can account for two thirds of the movements in the skill premium.[44]

3.4.2 Directed Technical Change

The theory of directed technical change maintains that a growth in the relative supply of a factor of production should induce technical change biased in favor of that factor. Historically, there are two important episodes of largely "exogenous" spurs in relative factor supply.

First, there was an increase in the supply of unskilled labor in urban areas of England during the 19th century. A careful look at the nature of technological progress over this period supports the theory. Goldin and Katz (1998) argue that in the 19th century the wave of technological innovations substituted physical capital and raw labor for skilled artisan workers (Braverman, 1974 and Cain and Paterson, 1986). For example, automobile production began in artisanal shops where the car was assembled from start to finish by a small group of "all-around mechanics." Only a few decades later, the Tayloristic model of manufacturing would bring together scores of unskilled workers in large-scale plants to assemble completely standardized parts in a fixed sequence of steps for mass production.

Second, there was a surge in skilled labor (i.e., workers with literacy and numerical skills) due to the "high-school movement" of 1910-1940. As pointed out by Aghion (2002), with respect to this episode, the theory finds weaker support. On the one hand, as we discussed earlier, it appears that the first part of the 20th century indeed marked the beginning of

[44]Note that this model implies that the origin of capital-skill complementarity is to be located only around 1950, later than what was argued in Goldin and Katz (1998).

a transformation in production methods biased towards skilled labor (from assembly lines to continuous and batch production processes). On the other hand, there was a decline in the returns to high school and the returns to college were stable (see Figure 4). Why is it that this wave of skill-biased technical change, which was as strong as the one 50 years later, did not have a similar impact on the wage structure? This question remains unanswered to date.[45]

3.5 Technology and the Gender Gap

Here we explore briefly the interaction between the gender gap and the advancements of technological change, both in the market and in the household.

3.5.1 Technological Change in the Market

As evident from Figure 1 (bottom panel), since the mid-1970s the gender wage gap has closed substantially. Several studies have concluded that this is due to a rise in relative labor demand for women, as supply cannot have played a large role (Bertola et al. 1997). Was the recent technological revolution "gender-biased"?

Consider a simple model where jobs differ in their requirement of physical effort and all jobs are necessary for production of the final good. At the same time, men and women have two traits: physical ability and cognitive ability. The theory of comparative advantage then implies that men will be most efficiently assigned to jobs with high physical requirements and that women should work on jobs with a large fraction of cognitive tasks.

The arrival of a new technology, like computers, that increases productivity relatively more on jobs with high cognitive content therefore tends to raise the average wage of women more than it raises the average wage of men. Weinberg (2003a) tests this theory on microeconomic data for the United States and finds that the increase in computer use for women can explain up to 50 percent of the increase in the relative demand for female employment.

It is worth noting that the gender premium fell in spite of the fact that the female-male

[45]Institutions might have played a role in the 1940s. Goldin and Margo (1992) argue that the National War Labor Board operated an explicit policy of wage compression during that period.

relative supply ratio grew almost by a factor of 2 between 1960 and 2000, i.e., by as much as the growth in the relative supply of college-educated labor. In the perspective of the directed technical change literature, one is left to ponder whether rising female participation was also a force that led innovators to spend resources on capital goods complementary with cognitive skills rather than with physical skills in order to exploit women's comparative advantage. This hypothesis remains to be analyzed in detail.

3.5.2 Technological Change in the Household

The postwar period witnessed another form of technological revolution: one that did not take place in factories and plants, but rather in the household. Greenwood et al. (2004) argue that the decline in prices of household appliances (refrigerators, vacuum cleaners, washers, dishwashers, etc.) worked as "engines of liberations" for women: new and more productive capital in the households could free up potential hours to be supplied in the labor market. In particular, as household durables were introduced into the economy, the effective wage-elasticity of female labor supply increases, which, in turn, helps explaining the sharp rise in female market participation, even in the presence of not-so-large changes in the gender wage gap.[46]

4 Technical Change and the Returns to Experience

According to Card and DiNardo (2002), one of the most important challenges to the hypothesis that the recent changes in the wage structure are linked to technological progress is to explain the combination of the rise in the returns to labor market experience for the low-educated workers in the population and the flat, or declining, pattern of the experience premium for college graduates.

It turns out that the existing theoretical literature does not provide a unified answer to the question of how technological change affects the experience premium. Examples of the literature we review in this section include job-specific or technology-specific experience that,

[46] We refer the reader to Greenwood and Seshadri (2004) in this Handbook for a detailed analysis of this channel.

in principle, may be adversely affected by technological change, but that may also benefit from technological change if that change is of a 'general purpose' variety, that is, if it makes experience more widely applicable.[47] We also look at general labor-market experience as a vehicle to lower the cost of adapting to technological change.

4.1 Experience with General Purpose Technologies

An important feature of the recent technological developments that has not received much attention in the literature on inequality is its *general purpose nature*. Aghion et al. (2002) formalize the idea of "generality" of a technology and build a theoretical framework to understand how it affects various dimensions of wage inequality, such as the experience premium. They model generality in relation to human capital: a more general technology allows a larger degree of *transferability* of sector-specific experience across the different sectors of the economy. For example, the ability to use computers for word-processing or programming is useful in numerous sectors and jobs in the economy.[48] Given that actual technological change is uneven across sectors, transferability of experience then increases the value of experience, that is, the experience premium.

Consider a simple overlapping-generations (OLG) model with two-period lived agents, and two production sectors indexed by $i = 0, 1$. Each cohort of agents has measure one and works in both periods. Technological progress results in capital-embodied innovations that increase productivity by a factor $1 + \gamma$ occurring in each of the two sectors in alternation. Let "0" denote the new sector in the current period. Suppose, for simplicity, that production takes place with a fixed amount of capital, normalized to one: the production function in sector i (in the stationary transformation of the model) is $y_i = A_i^{\alpha} h_i^{1-\alpha}$, where A_i measures

[47] We will return to the issue on how technological change interacts with the accumulation of job/technology-specific knowlege in the frictional models of Section 7.3.

[48] A survey conducted by the U.S. Bureau of Labor Statistics emphasizes that

> *"...the technology, network systems, and software is similar across firms and industries. This is in contrast to technological innovations in the past, which often affected specific occupations and industries (for example, machine tool automation only involved production jobs in manufacturing). Computer technology is versatile and affects many unrelated industries and almost every job category"* (McConnell, 1996, page 5).

the efficiency of capital in sector i ($A_0/A_1 = 1+\gamma$), and h_{it} measures the effective labor input in sector $i = 0, 1$.

Young agents are always productive on the new technology, whereas old workers can productively move to the new sector only with probability σ. This captures the idea that young workers are more "adaptable" than old workers, possibly because of vintage effects in their schooling, or because the ability to learn declines with age. Moreover, assume that this "adaptability constraint" is binding, in the sense that: (1) the equilibrium fraction σ^* of old workers who moves equals σ; and, (2) there is not enough labor mobility (σ is sufficiently low) to offset the impact on wages of the sectoral productivity differential $1+\gamma$.

Newborn agents start working in the new sector with initial knowledge normalized to 1.[49] Agents accumulate η additional units of experience through learning-by-doing in the first period of work. The generality of the technology determines the degree of skill transferability for the old workers, τ_o, i.e., the fraction of accumulated knowledge η a worker can carry over if she moves to the leading-edge sector at the beginning of her second period of life. The entire knowledge η can be used if the worker stays in the old sector.

Aggregate human capital in the old sector h_1 is determined by old, non-adaptable workers, a fraction $1-\sigma$, who have accumulated $1+\eta$ units of experience. Human capital in the new sector is determined by the new cohorts that have one unit of experience, and old adaptable workers with transferable experience, that is, $h_0 = 1+\sigma(1+\tau_o\eta)$. With competitive labor markets, the ratio between the prices of efficiency units of labor in the old and the new sector therefore is:

$$\frac{w_1}{w_0} = (1+\gamma)^{-\alpha}\left(\frac{h_0}{h_1}\right)^{\alpha} = (1+\gamma)^{-\alpha}\left[\frac{1+\sigma(1+\tau_o\eta)}{(1-\sigma)(1+\eta)}\right]^{\alpha}. \tag{15}$$

The steady-state experience premium, i.e. the average wage of old workers relative to the average wage of young workers, is therefore given by

$$x^* = \sigma(1+\tau_o\eta) + (1-\sigma)(1+\eta)\frac{w_1}{w_0}, \tag{16}$$

where one can see immediately that x^* is increasing in τ_o. That is, an increase in the generality of technological knowledge raises skill transferability and amplifies the experience

[49] Aghion et al. (2002) show that this is indeed the optimal choice of young cohorts, for general conditions.

premium of adaptable workers, who are able to transfer more of their cumulated skills. It also indirectly raises the experience premium of non-adaptable old workers by making effective adaptable labor input relatively more abundant in the economy, hence depressing the wage of young workers.

This result is particularly interesting in light of the fact that a version of this model that is based purely on the hypothesis that the rate of embodied technical change, γ, has accelerated would predict a decline in the experience premium. This is evident from the fact that the wage ratio w_1/w_0 is decreasing in γ: larger productivity differentials between the young and the old vintages represent a relative advantage to young workers who are more adaptable.

The more general model in Aghion et al. (2002) also features a flexible choice of capital. Another interpretation of generality of the technology offered in their paper is based on the *compatibility* of physical capital, i.e., the extent to which capital equipment embodying the old technology can be retooled—so as to embody the new leading-edge technology—and moved to the new sector. Under this interpretation, the arrival of a GPT, which increases the compatibility across vintages of capital, reduces the experience premium since it allows the transfer of more capital to the new sector where it benefits the young, inexperienced, but more adaptable workers.[50]

4.2 Vintage-Specificity of Experience

According to the GPT hypothesis, human capital becomes more transferable across sectors once the new technological platform has fully diffused throughout the economy. However, it is also reasonable to think that, at least in the transition phase, certain skills associated to the old way of producing quickly become obsolete. Or, put differently, human capital is vintage-specific. Thus, although in the final steady state skill transferability will be higher, it can undershoot during the transition.

[50]The model by Caselli (1999) outlined in section 3.2 has exactly this feature of capital mobility from the old technology to the new and more productive technology; thus, a version of that model where the young workers are those with the lowest learning cost would have the same counterfactual prediction for the experience premium.

To study the implications of vintage human capital for the experience premium, we can slightly modify the two-period OLG model in the previous section. To make this point starkly, consider the extreme case where old workers never find it profitable to move across sectors, so $\sigma^* = 0$, and suppose that when young workers join the new sector they lose a fraction $1-\tau_y$ of their initial knowledge (as before, normalized to 1). Modifying appropriately the equilibrium wage ratio (15) , equation (16) for the experience premium becomes

$$x^* = \frac{(\tau_y + \eta)\, w_1}{\tau_y w_0} = \frac{1}{(1+\gamma)^{\alpha}} \left[\frac{\tau_y + \eta}{\tau_y}\right]^{1-\alpha}, \tag{17}$$

which shows that x^* is decreasing in the skill transferability rate for young workers, τ_y. The arrival of a new technology that makes the knowledge of its (young) users obsolete can widen the returns to experience.

In analyzing earlier equation (16) we argued that a rise in γ would depress the experience premium, which is a problem for the pure "acceleration hypothesis". Vintage human capital can overturn this result. Suppose, as in Violante (2002), that the degree of skill transferability is decreasing in the speed of technological improvements, i.e. $\tau_y = (1+\gamma)^{-\tau}$. Then, it is easy to see from (17) that as long as $\tau > \alpha/(1-\alpha)$, the experience premium will rise after a technological acceleration, since the loss of vintage-specific human capital incurred by young workers is larger than the productivity improvement embodied in physical capital.[51] In Section 7.3 we return to the role of vintage human capital and discuss the plausibility of the assumption that the extent to which skills are transferable depends on γ.

4.3 Technology-Experience Complementarity in Adoption

According to the standard technology adoption models, the adopters of the new technology are likely to be the young workers because they face a lower learning cost or a longer time horizon to recoup the adoption costs. Weinberg (2003b) challenges this view and argues that there is one other force that gives more experienced workers an advantage: complementarity between new technologies and skills, together with the fact that more experienced workers are

[51] Note that this large skill loss for young workers does not necessarily imply that it is not optimal for them to begin working in the new sector. Indeed, by working with the new technology in the current period they improve *future* skill transferability.

more skilled, should lead to the prediction that older workers will adopt the new technology. What force dominates? And what are the implications for the experience premium?

Weinberg looks at the empirical pattern of computer usage (i.e., adoption of one of the new recent technologies) over the life-cycle and shows that it differs dramatically between high-school graduates and college graduates (see Figure 5).

FIGURE 5

Among uneducated individuals the profile is hump-shaped and peaks around 30 years of experience, while for educated individuals it is downward-sloping. As expected, the adoption rates for college graduates are higher at any given age.

These data suggest that the answer to the first question above depends on the level of schooling: for low-educated workers, experience is a substitute for general education, and the more experienced workers are also more productive in the new technology. Workers with high education levels are all equally adaptable to the new technology, so, for such workers, additional experience has a small marginal return in adoption. Since the learning cost increases with age, the youngest are more likely to adopt the new technology.

Adding to this mechanism the assumption that new technologies are more productive yields that the adopters gain a wage increase, which is consistent with the different pattern of the experience premium for low and high education groups that we described in Section 2.

4.4 On-the-Job Training with Skill-Biased Technological Change

The models reviewed in this section treat the degree of skill transferability or adaptability of workers as exogenous. If old workers recognize that "new knowledge" is necessary for dealing with the transformed technological environment, then one should expect that they would be willing to forego some resources to acquire such skills through training.

Mincer and Higuchi (1991) advanced this hypothesis and found some supporting evidence from U.S. sectoral data: industries with faster productivity growth were also the ones with steeper experience profiles and lower job-separation rates. They interpreted these facts as

reflecting the training channel in light of the findings of Lillard and Tan (1986) showing that the incidence of firm-specific on-the-job training is higher in sectors with high rates of productivity growth. Interestingly, Bartel and Sicherman (1998) document that the marginal impact of a rise in productivity growth on the likelihood of training (thus on the steepness of the wage profile) is stronger for low-educated workers, which is consistent with the pattern of the last 30 years mentioned in section 2.

The model developed by Heckman et al. (1998) explains the recent dynamics of the experience premium based precisely on this mechanism. To simplify the exposition, consider again a two-period OLG model where risk-neutral workers are endowed with a unit of human capital, work in both periods, and choose how much time to devote to on-the-job training and production in the first period. Training increases human capital in the second and final period. The problem of a worker at time t is:

$$\max_{\tau_t} \{w_t (1 - \tau_t) + \beta w_{t+1} h_{t+1}\}$$

$$s.t. \qquad h_{t+1} = \tfrac{A}{\theta} \tau_t^{\theta},$$

where τ_t is the fraction of the unitary time endowment spent in training, β is the discount factor, w_t is the wage rate at time t, and h_t is human capital at time t. We assume that production of human capital has decreasing returns in the time input. It is easy to see that optimal training and human capital are functions of expected wage growth:

$$\tau_t = \left(A\beta \frac{w_{t+1}}{w_t} \right)^{\frac{1}{1-\theta}} \text{ and } h_{t+1} = A^{\frac{1}{1-\theta}} \left(\beta \frac{w_{t+1}}{w_t} \right)^{\frac{\theta}{1-\theta}}.$$

The implied experience premium, that is, the wage of an experienced old worker relative to the wage of an inexperienced young worker at a given point in time, is then $x_t = h_t / (1 - \tau_t)$.

In a stationary state where $w_t = w^*$ for any t, the optimal fraction of time spent in training is $\tau^* = (A\beta)^{1/(1-\theta)}$, and the corresponding steady-state experience premium is

$$x^* = \frac{1}{\theta} \left(\frac{A^{\frac{1}{1-\theta}} \beta^{\frac{\theta}{1-\theta}}}{1 - (A\beta)^{\frac{1}{1-\theta}}} \right).$$

The steady-state experience premium is increasing in the productivity of training, A, and in the discount factor, β.

Suppose now that the economy undergoes a one-period transition toward a permanently higher level of skill-biased productivity. High-education (low-education) workers see their wage going up (down), i.e., $w_{t-1} = w_t = w^*$, $w_{t+n} = \bar{w}$ when $n > 1$, where for high-educated workers $\bar{w} > w^*$, and for low-educated workers $\bar{w} < w^*$. Since the two cases are perfectly symmetric, we solve for the transitional dynamics in the experience premium of the high-educated. Along the transition, in period t educated workers increase their investment in training since the anticipated rise in their wages increases the return to human capital accumulation, whereas in all future periods, i.e., $t+1$ and higher, educated workers do not change their human capital investment decision since their anticipated wage change is not affected:

$$\tau_t = \left(A\beta\frac{\bar{w}}{w^*}\right)^{\frac{1}{1-\theta}} > \tau^*, \tau_{t+n} = \tau^* \text{ for } n \geq 1.$$

The implied sequence of experience premia for educated workers is given by

$$x_t = \frac{h^*}{1-\tau_t} > x_{t+1} = \frac{h_{t+1}}{1-\tau^*} > x_{t+2} = x^*. \tag{18}$$

The experience premium first rises from x^* to x_t and then falls gradually towards the steady state. For low-educated workers, the opposite pattern will hold. If one thinks of time $t-1$ as 1965, i.e., the moment before the rise in inequality started, time t as 1975, and so on, this stylized model can qualitatively explain the rise in the experience premium for the less educated workers and the decline in the experience premium for the more educated in the 1980s.

The key force is the intertemporal substitution between working and training that the expected changes in wages bring along.[52] Also, as emphasized by Heckman et al. (1998), it is important to recognize that movements in earnings, $w(1-\tau)$, can differ from movements in skill prices w when labor supply is endogenous. The major limit of the theory is probably that the mechanism depends crucially on the ability of agents to perfectly foresee changes in wage rates decades in advance.

[52]Dooley and Gottschalk (1984) also explore a mechanism based on human capital investment in order to explain the rising inequality within cohorts of young workers in the United States. They attribute the changes in expected wages to aggregate fluctuations in labor force growth: the baby-boom and, subsequently, the baby-bust.

5 Inside the Firm: the Organization of Work

Hayek (1945) argued that a fundamental problem of societies is how to use optimally the knowledge that is available, but is dispersed across individuals. In frictionless markets, prices can solve this problem: they transmit knowledge about relative scarcity and relative productivity of resources. Since Coase (1937), it is well understood that frictions limit the efficiency of markets, and they divert certain transactions to occur within the boundaries of firms. Within the firm, the organization of work and production plays the role of the market as "information processor" to allow efficient use and transmission of knowledge.

It is therefore not surprising that the recent innovations that revolutionized the way in which information and communication takes place have affected the workplace organization within firms and the boundaries of firms. Their impact on the wage structure is perhaps less clear. The maintained hypothesis in the literature is that the recent episodes of reorganization of production, especially in manufacturing, have favored adaptable workers who have general skills and who are more versed at multi-tasking activities. An alternative view, which we will develop later in this section, is that organizational change is not induced by technological change, but that the increased relative supply of skilled labor created the incentives to change the organization of production.

5.1 The Milgrom-Roberts Hypothesis: IT-Driven Organizational Change

Milgrom and Roberts (1990) were the first to emphasize the interaction between the diffusion of information technologies in the workplace and the reorganization of production. Their hypothesis builds on the idea that information technologies reduce a set of costs within the firm which triggers the shift towards a new organizational design. First, electronic data transmission through networks of computers reduces the cost of collecting and communicating data, and computer-aided design and manufacturing reduces the costs of product design and development. Second, there are complementarities among a wide group of strongly integrated activities within the firm (product design, marketing, and production), and pronounced

non-convexities and indivisibilities in each activity.

As a result, as the marginal cost of IT declines, it is optimal to reorganize all activities to exploit this shock, and, due to non-convexities, organizational change can be sudden and drastic in nature. In particular, because of lower communication costs the layers in the hierarchical structure can be reduced, so that the organization of the firm becomes "flatter."[53] Workers no longer perform routinized, specialized tasks, but they are now responsible for a wide range of tasks within teams. These teams, in turn, communicate directly with managers. Because of the flexibility of IT capital, the scale of production decreases (recall the evidence in Mitchell (2001) on plant size), allowing greater production flexibility and product customization.

An elegant formalization of this hypothesis is contained in Bolton and Dewatripont (1994). They study the optimal hierarchical structure for an organization whose only objective is that of efficiently processing a continuous flow of information and show using their model that a reduction in communication costs leads to a flatter and smaller organization.

5.1.1 Implications for the Wage Structure

Although in their original papers neither Milgrom and Roberts nor Bolton and Dewatripont explore the implications of organizational change for the wage structure, a small but growing literature on IT-driven organizational change and inequality has developed since.

Lindbeck and Snower (1996) emphasize the "complementarity" aspect of the Milgrom-Roberts hypothesis. They consider a production function with two tasks and two types of workers. The Tayloristic model would assign one type of workers to each task, according to comparative advantages to exploit specialization. The alternative organization of production is the flexible model, where each type of worker performs both tasks. This more flexible organization is preferred when there are large informational complementarities across tasks. The introduction of IT capital amplifies these informational complementarities and makes the flexible organization more profitable. Moreover, firms increase the demand for skilled

[53]Rajan and Wulf (2003) use detailed data on job descriptions in over 300 large U.S. companies to document that the number of layers between the lowest manager and the CEO has gone down over time, i.e., organizations have become "flatter".

workers who are more adaptable and versed in multi-tasking, and the skill premium rises.

Möbius (2000) focuses on the "customization" aspect of organizational change. When products are standardized, demand is certain and production tasks perfectly predictable, inducing a high division of labor (the Tayloristic principle). New flexible capital allows firms to greatly expand the degree of product variety and customization in product markets. Larger variety implies a more uncertain demand mix because producers become subject to unpredictable "fad shocks" and producers therefore favor a flexible organization of production, with less division of labor. Once again, to the extent that the most skilled workers are also the most adaptable and versatile, the skill premium will increase.

The mechanism in Garicano and Rossi-Hansberg (2003) is based, instead, on the fall in the communication cost within the organization. Their paper has the particular merit of taking the literature on the internal organization of firms (e.g., Bolton and Dewatripont 1994) one step further by recognizing that organizational hierarchies and labor market outcomes are determined simultaneously in equilibrium. Consider an organization where managers perform the most difficult and productive tasks and workers specialize in a set of simpler tasks. Managers also spend a fraction of their time "helping" workers unable to perform their task, and by so doing, they divert resources away from their most productive activities. The fall in the cost of communication allows workers to perform a wider range of tasks, using a smaller amount of the manager's time. The implications for wage inequality are stark. First of all, since workers are heterogeneous in ability, and ability is complementary to the number of tasks performed, inequality among workers within the firm increases. Second, the pay of the manager relative to that of the workers rises because the manager can concentrate on the tasks with high return.

The previous papers have studied how IT-based advances have affected the organizational structure within firms. Saint-Paul (2001) addresses the spectacular rise in the pay of CEOs and a few other professions (e.g., sportsmen and performers) documented in Section 2 using a model where IT-based advances affect the organization of markets with frictions. Saint-Paul combines a model with "superstar" or "winner-take-all" effects (Rosen 1981) with the advent

of information technology. In his model, human capital has two dimensions: productivity, i.e., the ability to produce units of output, and creativity, i.e., the ability to generate ideas that can spread (and generate return) over a segment of an economy, called a "network." The diffusion of information technology expands networks increasing the payoff to the most creative workers and widening the income distribution at the top. However, as networks become large enough, the probability that within the same network there will be somebody with another idea at least as good rises: superstars end up competing against each other, mitigating the inegalitarian effects of information technology. Under certain parametric assumptions, inequality first rises and then falls over time.

5.1.2 Empirical Evidence on the Complementarity between Technology, Organizational Change and Human Capital

Bresnahan et al. (2002) investigate the hypothesis that IT adoption, workplace reorganization, and product variety expansion (customization) are complementary at the firm level. Their view is that simply installing computers or communications equipment is not sufficient for achieving efficiency gains. Instead, firms must go through a process of organizational redesign. The combination of IT investments and reorganization represents a skill-biased force increasing the relative demand for more educated labor.

Their empirical analysis is based on a sample of over 300 large firms in the United States, and their definition of organizational change is a shift towards more decentralized decision making and more frequent teamwork. They find a significant correlation between IT, reorganization, and various measures of human capital.[54]

In a related paper, Caroli and Van Reenen (2001) argue that the existence of complementarities between organizational change and the demand for skilled labor leads to three predictions: 1) organizational change should be followed by a declining demand for less skilled workers; 2) in the vein of the directed technical change hypothesis (see next section), cheaper skilled labor should increase the occurrence of organizational change; and 3) organizational change should have a larger impact in workplaces with higher skill levels.

[54]See Brynjolfsson and Hitt (2000) for a survey on the empirical work documenting the causal link from adoption of information technology and organizational transformation within the firm.

They test these predictions combining two data sets, one for the United Kingdom and one for France, with information on changes in work organization, working practices, and the skill level of the labor force. Interestingly, they also have information on the introduction of new IT capital, so they can distinguish the effect of organizational change from that of skill-biased technical progress. They find some supporting evidence for all three predictions.

Baker and Hubbard (2003) offer an example where technological change not only affects the organizational design of firms but also the boundary of firms. In particular, they study how IT may have reduced the moral hazard problem in the U.S. trucking industry. Drivers may simply operate the trucks as employees of the dispatching company, or they may actually own the trucks they operate. If the dispatcher owns the truck, there is only limited assurance that the driver will operate in a way that preserves the value of the asset, since the dispatcher cannot perfectly monitor the driving operations. When this moral hazard problem is severe, decentralized ownership will be the outcome, that is, the driver owns the truck. Using detailed truck-level data, Baker and Hubbard show that with the introduction of a new monitoring technology—on-board computers linked to the company servers—the share of driver-ownership decreased significantly.

5.2 Directed Organizational Change

An alternative hypothesis to that put forth by Milgrom and Roberts is contained in several papers discussing the parallel change in the organization and in the pay structure of work. This view maintains that the driving force of organizational shifts is not technology, but rather the secular rise in the supply of skilled workers that created incentives to modify the organization of production: directed organizational change of sorts.

Acemoglu (1999) models a frictional labor market where firms must choose the amount of capital, k, when they are vacant, before meeting the worker. Consider a simple static version of Acemoglu's model. There are two types of workers, skilled and unskilled, where ϕ is the fraction of skilled ones. Skilled workers have productivity, h_s, and unskilled workers, h_u, which we normalize to $1 < h_s$. Output on each job is given by $y_i = h_i^{\alpha} k^{1-\alpha}$, where $i = s, u$. Wages and profits and are, respectively, a fraction ξ and $1 - \xi$ of output net of the cost of

the capital installed k. The expected value of a firm choosing capacity k is

$$V(k) = (1-\xi)\left[\phi I^s\left(h_s^\alpha k^{1-\alpha} - k\right) + (1-\phi) I^u\left(k^{1-\alpha} - k\right)\right],$$

where I^i is an indicator variable that equals 1 if the firm accepts a match with a worker of type $i = s, u$ and 0 otherwise.[55] Suppose the firm chooses between two hiring strategies: a "pooling" strategy where it accepts all workers, $I^s = I^u = 1$, and a "separating" strategy where it only accepts skilled workers, $I^s = 1$, $I^u = 0$. Conditional on the hiring strategy, we can use the first-order condition to solve for the optimal choices of capacity, k^P and k^S. Substituting the capacity choice back into $V(k)$, the values of the two hiring strategies are:

$$\begin{aligned} V^P &= \kappa(1-\xi)\left[\phi h_s^\alpha + (1-\phi)\right]^{1/\alpha}, \\ V^S &= \kappa(1-\xi)\phi h_s, \end{aligned} \tag{19}$$

where κ is a constant depending only on α. Comparing these two values, we conclude that the payoff to the "separating" strategy, V^S, dominates the payoff of the "pooling" strategy, V^P, whenever

$$\left(\frac{1-\phi}{\phi^\alpha - \phi}\right)^{1/\alpha} < h_s. \tag{20}$$

Note that the left-hand side of this expression decreases in ϕ, the fraction of skilled workers. When the size of the skilled group is small, a "pooling" equilibrium arises where all firms invest the same amount of capital and search for both types of workers. As the relative size of the skilled group rises, the economy switches to a "separating" equilibrium where firms find it optimal to install more capital and accept exclusively skilled workers in their search process.[56] One can interpret the pooling and the separating equilibrium as different types of work organizations, displaying different degrees of segregation along the skill dimension within sectors. The switch from the low-segregation to the high-segregation organization stretches the wage structure and generates higher inequality.

In a related paper, Kremer and Maskin (1996) offer an alternative explanation for the rise in the degree of assortative matching in the workplace, using a frictionless assignment model.

[55] Here, for simplicity we assume that workers accept passively each job offer. We do not consider equilibria where firms randomize, i.e., where $I^i \in (0,1)$.

[56] In the more general version of the model, which is dynamic with free entry of firms, there are other firms who install a small amount of capital (unskilled jobs) and search exclusively for unskilled workers.

Their paper contains some suggestive evidence that the degree of sorting ("segregation") has risen within industries and plants. However, their model is based on an increase in the skill dispersion in the population, for which there is little evidence in the data.[57]

Thesmar and Thoenig (2000) embed a choice of organizational design into a Schumpeterian growth model. Firms can opt for a Tayloristic organization that has large product-specific set-up costs, with the benefit of a high level of productive efficiency. Alternatively, they can choose a new and more flexible organization that can be built with a lower initial fixed cost, but whose productivity level is lower.[58] As is common in this class of Schumpeterian models, there is an R&D sector, where product innovations are generated proportionately to the amount of skilled workers hired. The patent of each new product is then sold to a monopolistic producer who can choose optimally which organization of work to set up (Tayloristic or flexible) according to the volatility of the economic environment.

A rise in the supply of skilled workers will increase the innovation rate in the R&D sector: the higher the innovation rate, the shorter the product's life expectancy for a monopolistic producer, and the less profitable organizations with large fixed costs prove to be, compared to the more flexible production method. The model also produces a rise in segregation, since skilled workers tend to cluster into the R&D sector, as well as a rise in inequality as unskilled workers lose from the abandonment of the Tayloristic model since the production phase becomes less efficient.[59]

5.3 Discussion

The case examined by Baker and Hubbard (2003) is one where IT improves firms' monitoring ability of workers' effort. However, it is plausible that the trend towards a "flatter" organizational design where single-task routinized work is replaced by multi-tasking team-

[57]For example, Hoxby and Long (1998) report that the difference in the quality of education (measured by their wage) received by college students at more and less selective institutions has increased over time, but the increase is quantitatively small.

[58]This distinction between the Tayloristic firm and the new flexible firm is due to Piore and Sabel (1984).

[59]Duranton (2004) provides yet another framework for formalizing the concept of "skill segregation" in production and analyzes the implied wage structure in the economy. In his model, a rise in the relative supply of skilled workers can lead to higher segregation and more inequality.

work induces a *rise* in the cost of monitoring individual workers' effort. Firms would then, optimally, introduce incentive schemes (e.g., tournament contracts) with the result of increasing inequality in rewards. In other words, optimal contracts respond to technological and organizational changes that affect the extent of moral hazard within the firm. This line of research is largely unexplored at the moment.

All the models we surveyed in this section are qualitative in nature and, although they establish a logical link between organizational change and inequality, they do not provide any quantitative analysis. One of the main obstacles is that explicit models of organizations contain parameters and variables that are hard to observe, measure, and therefore calibrate (hierarchies, communication costs, number of tasks, etc.). Recently, several papers have started to measure, in various ways, "organizational capital" or "intangible capital" (see, e.g., Hall, 2001, McGrattan and Prescott, 2003, and Atkeson and Kehoe, 2002). A promising avenue for research would try to incorporate this measurement into models that link reorganization with changes in the stock of organizational capital and that relate the latter to the wage structure in order to perform a more rigorous quantitative analysis.

6 Technical Progress as a Source of Change in Labor Market Institutions

Throughout the chapter, up to this point, we have maintained a "competitive" view of the labor market and argued that skills are priced at their marginal product, potentially explaining large parts of the observed dynamics of inequality. However, the labor market displays very peculiar features compared to many other markets in the economy: a sizeable fraction of labor may be considered as under-employed in any given period (unemployment), individual workers often organize themselves into coalitions (unions), and wages frequently seem to be set through some explicit negotiation between firms and workers (individual and collective bargaining). These attributes of the labor market are, arguably, better captured by non-competitive models. We begin our departure from the purely competitive framework by introducing unions and collective bargaining.

Historically, unions and centralized bargaining have been key institutions in the determination of wages and other important labor market outcomes. Over the past 30 years, the economies of the United States and the United Kingdom experienced rapid deunionization. In the United States, in the late 1970s, 30 percent of male non-agricultural private-sector workers were unionized. By 2000, only 14 percent were unionized (Farber and Western, 2000). In the United Kingdom, union density among male workers was around 58 percent in the late 1970s and it has fallen uninterruptedly since to 30 percent today (Machin, 2000 and 2003). There is a variety of evidence that unions compress the structure of wages, even after controlling for workers' characteristics, and thus many economists suspect that their decline may have been an important factor in the increase in inequality in the Anglo-Saxon economies (see, e.g., Gosling and Machin, 1995, and DiNardo et al. 1996).

The existing literature has explored mainly two explanations for the decline in unions. The first generation of papers argued that an important force in the fall of unionization is the change in the composition of the economy away from industries, demographic groups, and occupations where union organization was comparatively cheaper and unions have been traditionally strong (Dickens and Leonard, 1985). However, Farber and Krueger (1992) estimate that compositional shifts can account for at most 25 percent of the decline in the United States and have played virtually no role since the 1980s. Machin (2003) reports that only around 20 percent of the U.K. union decline of the last two decades can be attributed to compositional change.

The second hypothesis is that the legal and political framework supporting union membership deteriorated in the 1970s and 1980s.[60] To date, this explanation seems to have gained rather broad acceptance, even though this view has limits as well. For example, the fall in union organizing activity precedes two key political events: the air-traffic controller strike of 1981 and Reagan's Labor Board appointments in 1983 (Farber and Western 2002). U.K. data also show that the fall in union membership pre-dates the first Thatcher government. Overall, we think that the forces behind rapid deunionization are not yet well understood.

[60]Some authors emphasized anti-union management practices (Freeman 1988). Others focused on changes in the composition of the National Labor Relation Board (Levy, 1985).

In most of continental Europe, unions are still strong, and there are no clear signs of decline in union coverage, but a marked change in union behavior has occurred over the past 30 years. Several indexes of coordination and centralization in unions' bargaining for Europe show a distinct trend towards more decentralized wage negotiations, especially in the Scandinavian countries, whose unionization rates are the highest (Iversen, 1998).

The standard explanation for the shift towards decentralized bargaining is based on the interaction between monetary policy and wage setting arrangements. With an independent national central bank, coordination in bargaining among unions is useful because it allows unions to internalize the implications of their wage claims on inflation. With the advent of the European monetary union and the institution of the European Central Bank within-country coordination proves less useful. However, the evidence in favor of this hypothesis is scant. First, monetary policy does not seem to Granger-cause centralization empirically (Bleaney, 1996). Second, we did not observe a substantial trend towards cross-border coordination in unions' bargaining.

Recently, a new hypothesis for deunionization and decentralization in unions' wage setting, based on skill-biased technological change, has been advanced by Acemoglu et al. (2001) and Ortigueira (2002). Their arguments rest on the view that unions are coalitions of heterogeneous workers which extract rents from employers and only exist insofar as members have an incentive to stay in the coalition and continue bargaining in a centralized fashion. The conjecture of these authors is that skill-biased technical change can dramatically alter such incentives.

6.1 Skill-Biased Technology and the Fall in Union Density

Here, we outline a reduced form model that conveys the basic trade-offs highlighted by Acemoglu et al. (2001). Suppose there are two kinds of workers l_s of which are skilled and $l_u = 1 - l_s$ of which are unskilled. If employed in the competitive sector, these workers will receive wages equal to their productivity, h_s and $h_u < h_s$, respectively. We will think of skill-biased technological change as a rise in h_s relative to h_u.

Workers can also be employed in unionized firms and receive wages, w_s and w_u. A main

characteristic of unions is that they compress wages. In our setup, this means that the wage gap between the unionized skilled and unskilled workers is smaller than the productivity gap, or

$$w_s - w_u = \kappa (h_s - h_u), \tag{21}$$

where $\kappa < 1$ is the degree of wage compression. This equation may arise for a variety of reasons. Collective decision-making within a union may reflect the preferences of its median voter, and if this median voter is an unskilled worker, he will try to increase unskilled wages at the expense of skilled wages. It is also possible that union members choose to compress wages because of ideological reasons or for social cohesion purposes. Or, in presence of idiosyncratic uncertainty, unions could offer insurance to their members by setting a flatter income profile. The empirical literature is broadly consistent with the notion that unions compress wages, though it does not distinguish among the various possible reasons for it (see Booth, 1995).

Union wages (w_s, w_u) must also satisfy some participation constraint for firms (who would otherwise either shut down or open a non-unionized plant). Suppose that this takes the form of non-negative profits:

$$h_s l_s + h_u (1 - l_s) + \Omega (h_s, h_u) - [w_s l_s + w_u (1 - l_s)] \geq 0, \tag{22}$$

where $\Omega (h_s, h_u) > 0$ is the additional contribution of unions to output, as a function of both types of labor.[61] This could be because unions, ceteris paribus, increase productivity (for example, Freeman and Medoff, 1984, and Freeman and Lazear, 1995, argue this). Or unions may encourage training (as in Acemoglu and Pischke, 1999).

Solving the wage compression and participation constraint equations (21) and (22) as equalities, we obtain the maximum wage that a skilled worker can be paid as a union member:

$$\overline{w}_s = h_s - (1 - l_s)(1 - \kappa)(h_s - h_u) + \Omega (h_s, h_u).$$

Intuitively, as w_s rises, w_u must increase too in order to satisfy the wage compression constraint (21) but since profits fall with labor costs there is an upper bound to the wage of a

[61] As long as unions are sustainable, all workers, skilled and unskilled, will prefer to join the union.

skilled union member. Skilled workers will remain union members as long as what they are paid as union members exceeds their competitive salary,

$$\overline{w}_s \geq h_s. \tag{23}$$

From the no-quit condition (23) and the wage compression constraint (21), it follows that $\overline{w}_u \geq h_u$, so unskilled workers will always remain unionized. Observe that the slope of the maximum union skilled wage $\overline{w}_s$ as a function of the productivity of skilled workers h_s is:

$$\overline{w}'_s(h_s) = 1 - (1 - l_s)(1 - \kappa) + \Omega_1(h_s, h_u),$$

Since $\kappa < 1$, as long as the benefits of unionization, $\Omega(h_s, h_u)$, do not increase too rapidly in h_s (i.e., the benefits of unionization do not increase much with skill-biased technical change), we have $\overline{w}'_s(h_s) < 1$. Hence, there exists a cutoff level, h_s^*, such that $\overline{w}_s(h_s) < h_s$ for any $h_s > h_s^*$. This implies that once technical change takes h_s above h_s^*, the wage compression imposed by unions becomes unsustainable, and skilled workers will break away from unions.

Notice that skill-biased technical change is the cause of the deunionization and directly increases inequality. However, deunionization itself contributes to inequality as well. Before deunionization, the wage gap between skilled and unskilled workers is $w_s - w_u \leq \kappa(h_s - h_u)$, and widens smoothly with skill-biased technical change. It is only after deunionization that it jumps up discretely to $w_s - w_u = h_s - h_u$. Therefore, although deunionization is not the primary cause of the surge in wage inequality, it amplifies the original effect of these economic forces by removing the wage compression constraint imposed by unions.

6.2 Skill-Biased Technology and the Fall in Centralized Bargaining

In many European countries—in particular among the Scandinavian countries—the so-called "Ghent system" creates a fiscal-policy link among unions. Under this system, unemployment benefits are administered by the individual unions, but they are funded by the government through aggregate labor income taxation. Hence, not only does the net income of unions' members depend on their negotiated wage, but, through the equilibrium tax rate, also on

the wage claims of other unions. Ortigueira (2002) outlines a model economy with this institutional feature, where there are two types of workers, skilled and unskilled, and two unions that can choose to coordinate their wage determination. Unemployment is generated through a frictional labor market with a standard matching function.

Under decentralized bargaining, unions take the tax as given. Ortigueira (2002) shows that there are two possible steady states: in one, unions expect a low tax, thus making moderate wage claims which, in turn, keep equilibrium unemployment and tax rate low, fulfilling the initial expectation; in the other steady state, unions expect a high tax rate, thus making strong wage claims that produce high unemployment and a high tax rate. This second equilibrium yields lower income and lower welfare for union members. Centralized bargaining avoids the coordination failure and the associated welfare losses that can arise in this "bad equilibrium," and hence it can be preferred by unions. Note, however, that the "good equilibrium" under decentralized bargaining is still the best outcome. It is the ex-ante uncertainty that the bad equilibrium could arise that makes coordination attractive.

However, consider what happens with the advent of a skill-biased technology that increases the demand for skilled workers sharply, reducing their unemployment incidence. When unemployment benefits are proportional to wages, the fact that skilled workers are much less likely to be unemployed decreases the social expenditures of the government. As a result, under decentralized bargaining, the equilibrium with high taxes and low welfare does not survive the advent of a skill-biased technology. This justifies the shift in unions' wage setting policies towards decentralization.[62]

6.3 Discussion

The testable implications that can be identified above are that (1) among the experienced workers, the most skilled leave the unions in response to technological improvements and that (2) among the new entrant cohorts, the most educated workers opt for non-unionized

[62]See also den Haan (2003) for a model with multiple steady states, one with low tax and unemployment rates and one with high tax and unemployment rates, applied to the U.S.-Europe comparison of labor market outcomes.

jobs. However, these implications are derived from theories of technology-induced deunionization that are rather exploratory; more sophisticated and rigorous models of unions (with endogenous membership and endogenous wage-compression mechanisms) are yet to be developed.

The recent empirical studies by Card (2001), for the United States, and Addison et al. (2004), for the United Kingdom, compare the unionization rate across several skill groups before and after the collapse in union density in these two countries (1973 and 1993 for the United States, and 1983 and 1995 for the United Kingdom). The common finding of these two papers, is that unionization declined most for the low- and middle-skill groups.[63] Taken at face value, this preliminary evidence is not favorable to the hypothesis discussed in this section. However, one has to be cautious in interpreting these results because this work does not control for unobserved heterogeneity.[64] Suppose that—as documented by Card (1996)—unobserved ability is higher among unionized workers with low observable skills. Given that unionized firms offer a compressed wage schedule, such a contract would attract the highest ability workers with low education and the lowest ability workers with high education. Moreover, assume that technological change induces a rise in the market return for innate ability, as discussed in section 3.2. Then, the theory suggests that one should observe exactly the cross-skill deunionization pattern documented from U.S. and U.K. data.

It should be mentioned that a technology-based theory of deunionization must also explain why union density did not fall (in fact, it expanded somewhat) in the public sector. Since the public sector is, by definition, sheltered from the international competition, it is reasonable to conjecture that the leap in competitive pressure faced by many manufacturing industries over the past 30 years eroded those rents that are, according to some researchers, at the heart of the existence of unions. A quantitative evaluation of the importance of this channel is yet to be performed.

[63]Note that wages in the union sector do not fully reflect skills. For this reason, these authors impute skill deciles to unionized workers based on what workers with similar observable characteristics (age, education, gender, race, etc.) would earn in the non-union sector.

[64]Card (2001) makes a rough adjustment for unobserved heterogeneity, based on Card (1996). A thorough analysis would require the use of longitudinal data, but both, Card (2001) and Addison et al. (2004), are restricted to repeated cross-sections.

Another avenue that so far has not been pursued is the analysis of deunionization in conjunction with the structural changes in workplace organization that occurred in the past 30 years. In Section 5, we argued that a distinct feature of the recent change in the production process, especially in manufacturing, is the switch from Tayloristic organizations, where workers repeatedly performed similar tasks around the conveyer belt, towards "flatter" organization built on teams where workers engage in multiple tasks and where the individual division of labor is much fuzzier. Union's wage setting arrangements, based on "equal pay for equal work", can be effective within a Tayloristic plant, but then become very inefficient in plants where production is organized through teams. There is no reason to assume that workers performing the same task will be equally productive, since they perform many other complementary operations simultaneously (see, e.g., Lindbeck and Snower, 1996).

7 Technological Change in Frictional Labor Markets

Most of the models presented so far feature an aggregate production technology, i.e., the production structure is centralized, and competitive labor markets. Constructing a frictional model of the labor market requires departing from both attributes and moving towards a decentralized production structure and a labor market with imperfect coordination between workers and firms in the matching process. This class of models gives rise to frictional equilibrium unemployment and "frictional equilibrium inequality". By frictional inequality, we mean wage dispersion that is purely an artifact of frictions and that, without frictions, would disappear. A useful way to think about this phenomenon is to introduce the concept of "return to labor market luck".

Throughout this chapter, we have discussed several models where technological progress produces a rise in the return to observable and unobservable *permanent* components of individual skills, such as educational attainment, age, and innate ability. These permanent factors greatly determine inequality of earnings among the population, but they are not by any means exhaustive. Earnings display a large *stochastic* component (e.g., events related to the luck of individuals, firms, or industries) that is responsible for their fluctuations around

the permanent component.[65]

Gottschalk and Moffitt (1994) were the first to ask how much of the observed increase in inequality is attributable to a rise in earnings volatility and instability around its permanent component. They used a simple statistical model where log wages, w_{it}, for an individual i at time t—net of their predictable age profile—are assumed to be the sum of two orthogonal components, a fixed individual effect, α_i, and a stochastic ($i.i.d$) component, ε_{it}. Using the covariance structure of wages within a panel of U.S. males (constructed from PSID data), they reached the conclusion that the fraction of the total increase in cross-sectional inequality attributable to a surge in earnings volatility is between one third and one half.[66] One can interpret this fact as a rise in frictional inequality, or in the "return to labor market luck." The argument set forth is that the rapid diffusion of a new technology leverages the importance of these stochastic factors, raising the premium to workers with no observable distinguishing characteristics other than their good fortune.

Most of the work we review uses the random matching model of the labor market (see, e.g., Mortensen and Pissarides, 1998, or Pissarides, 2000). In this framework the existence of frictions creates a bilateral monopoly as a result of a meeting between a vacant firm and a worker. Wages are determined by bargaining over total output, so more productive firms tend to pay more, creating wage dispersion among ex-ante equal workers. We start by studying how technological change affects unemployment in this class of models. Next, we move to wage inequality. Random matching is a somewhat extreme characterization of frictions. In the last part of the section we contrast random search models to directed search models.

[65] A large empirical literature documents wage dispersion among observationally equivalent workers that cannot be fully reconciled with unobserved heterogeneity in permanent components. Abowd et al. (1999) document that firm effects still play a role, after controlling exhaustively for individuals' effects. Krueger and Summers (1988) found that a worker moving from a high to a low wage industry is subject to a wage loss roughly equal to the inter-industry differential.

[66] The subsequent literature on the subject demonstrated the robustness of this result to richer statistical models for the stochastic component of wages. See Haider (2001), Heathcote et al. (2003), and Meghir and Pistaferri (2004) for the United States and Blundell and Preston (1998) and Dickens (2000) for the United Kingdom.

7.1 Technological Progress and Frictional Unemployment

There is a sizeable literature trying to characterize how equilibrium unemployment reacts qualitatively to variations of the rate of technological change within a matching model à la Diamond-Mortensen-Pissarides (DMP) with vintage capital à la Solow (1960). Two distinct approaches emerge from the literature.

The first, that can be attributed to Aghion and Howitt (1994), argues that when new and more productive equipment enters the economy exclusively through the creation of new matches—because existing matches cannot be "upgraded"—it has a Schumpeterian "creative-destruction" effect: new capital competes with old capital by making it more obsolete and tends to destroy existing matches, because workers are better off separating from their old matches to search for the new firms endowed with the most productive technology. Thus, unemployment tends to go up as growth accelerates, due to a higher job-separation rate.

The second approach, due to Mortensen and Pissarides (1998), proposes an alternative view whereby the new technologies enter into existing firms through a costly "upgrading" process of old capital. In the extreme case where upgrading is free, we have the Solowian model of disembodied technological change, even though the carrier of technology is equipment. The separation rate is unaffected by faster growth and all the effects work through job creation. For small values of the upgrading cost, unemployment falls with faster growth, thanks to the familiar "capitalization effect": investors are encouraged to create more vacancies, knowing that they will be able to incorporate (and hence benefit from) future technological advances at low cost.[67]

Hornstein et al. (2003b) try to resolve the issue quantitatively. When they parameterize the model to match some salient features of the U.S. economy, they find that, in the vintage-matching model, the link between capital-embodied growth and unemployment does not

[67] An interesting qualification to this result is provided by King and Welling (1995): if, unlike what is customarily assumed in this family of models, workers bear the full fixed search cost, then the capitalization effect leads to an increase in the number of searchers and to longer unemployment durations. See Pissarides (2000, chapter 3) for a detailed discussion on growth and unemployment in matching models of the labor market.

importantly depend on to what parties—new matches or old ones—the benefits of the technological advancement accrue. The intuition for this "equivalence result" is that upgrading can be much better than creative destruction only if it is very costly for vacant firms to meet workers, but the data on the low average unemployment and vacancy durations imply that, in the model, this meeting friction is minor. That paper also shows that the same data on average unemployment duration impose severe restrictions on how much frictional wage inequality the model can generate. In the standard search model, high dispersion of wage opportunities makes workers very demanding and increases unemployment spells. Thus, a high wage dispersion could only coexist, in equilibrium, with long unemployment durations.

We now turn to the analysis of how technological progress impinges on frictional inequality in random matching models. In these models, however, the limits on the extent of wage inequality due to luck emphasized in Hornstein et al. (2003b) apply as well.

7.2 Technological Heterogeneity and the Returns to Luck

In a frictional labor market populated by ex-ante equal workers, an increase in technological heterogeneity can increase the return to luck. We explain this mechanism within a simple framework based on Aghion et al. (2002).[68] Consider an economy populated by a measure one of infinitely lived, ex-ante equal, and risk-neutral workers as well as by the same measure of jobs. Jobs are machines embodying a given technology. The technological frontier advances every period at rate $\gamma > 0$. The machines have a productive life of two periods. An age $j \in \{0, 1\}$ machine that is matched with a worker produces output, $y_j = (1+\gamma)^{-j}h$ (normalized relative to the age 0 machine), where h represents the skill level of the workers.

The labor market is frictional, i.e., workers separated from their jobs are randomly rematched with a vacant machine. To simplify, we assume that they always make contact with a machine. We postulate that, upon contact, the bilateral monopoly problem is solved by a rent sharing mechanism setting wages to be a constant fraction, ξ, of current output, y_j, where ξ is a measure of the bargaining power of workers.

It is easy to see that in an equilibrium where all job offers are accepted, the lucky half

[68] See also Manuelli (2000) and Violante (2002).

of the workers will be employed on new machines and the unlucky half on old machines. The variance of log wages is simply given by $var(\log w) = \gamma^2/4$, which is increasing in the rate of embodied technological change. Intuitively, in this economy all the heterogeneity is generated by technological differentials across machines. A technological acceleration (rise in γ) amplifies the productivity gaps between jobs. Since in this non-competitive labor market individual wages are linked to individual output, this acceleration then also raises wage dispersion even among ex-ante equal workers, i.e., it raises the return to luck.[69] As in Jovanovic (1998), however, if the scrapping age of capital is endogenous, the model would display an offsetting force. This force is due to the fact that, when the growth rate is higher, machines becomes obsolete faster, and firms scrap machines earlier. Therefore the equilibrium age range of machines in operation shrinks, compressing technological heterogeneity.

7.3 Vintage Human Capital with Frictions

A technological acceleration not only affects transitory residual wage inequality through its impact on the underlying distribution of job productivity differences. The technological acceleration may also affect the distribution of worker productivity differences if it interacts with the accumulation of job/technology-specific knowledge.[70] Violante (2002) extends the above model to include vintage human capital. Employed workers accumulate, through learning-by-doing, knowledge about the technology they are matched with. We normalize the amount of specific skills cumulated after every employment period to 1, so that the learning curve of the workers is concave, i.e., learning is faster for workers with lower initial skills. To keep the model tractable, we also assume that skills fully depreciate after two periods.

A worker on a machine of age i who moves on to a machine of age j next period can

[69] This increase in wage inequality is mirrored by a rise in wage instability along the lifetime of each worker: given a certain amount of labor turnover, larger cross-sectional productivity differences translate into more volatile individual wage profiles.

[70] The accumulation of job/technology-specific knowledge is also at the heart of the discussion of the experience premium in section 4.

transfer h_{ij} units of the accumulated skills to the new job :

$$h_{ij} = \min\left\{(1+\gamma)^{\tau(j-i+1)}, 1\right\}, \tag{24}$$

with $\tau > 0$ and $i, j \in \{0, 1\}$. The fraction of skills that can be transferred from an old to a newer machine is proportional to the *technological distance* between the two machines through a factor $\tau \geq 0$. The presence of the term γ in the transferability technology is crucial: the rate of quality improvement of capital-embodied technologies determines the degree by which new technology is different, more complex, and richer than the previous generation of machines. A higher γ reduces skill transferability in the economy.[71] Equation (24) and the depreciation assumption implies that we have three skill levels in the economy:

$$h_{01} = 1, \quad h_{00} = h_{11} = (1+\gamma)^{-\tau}, \quad h_{10} = (1+\gamma)^{-2\tau}, \tag{25}$$

and the corresponding wage rates (normalized relative to the wage on an age 0 machine) are $w_{ij} = \xi h_{ij} (1+\gamma)^{-j}$, $i, j \in \{0, 1\}$. Note that, given this simple expression for wages, the variance of log wages can be written as

$$var(\widetilde{w}) = \gamma^2 var(j) + var(\widetilde{h}) - 2\gamma cov(\widetilde{h}, j), \tag{26}$$

which is the sum of technological heterogeneity (the force discussed earlier), ex-post skill heterogeneity among workers, and the degree of assortative matching between skills and technologies measured by their covariance.[72]

One can prove that, for large enough γ, workers separate from firms every period.[73]

[71] The book by Gordon (1990) provides several examples of quality improvement in equipment requiring the performance of new tasks in the associated jobs. In the aircraft industry in the 1970s, new avionics were introduced that provided a safer but more complex navigation system. In the telephone industry, around the mid-1970s, electromechanical telephone switchboards were replaced by more sophisticated and flexible electronic equipment with larger programming possibilities. In the software industry, since the early 1980s, every new version of a software is equipped with new features. Those users who remain attached to an old version are often unfamiliar with many features of the new version.

[72] A rise in the degree of assortative assignment between workers' skill and machines' productivity is equivalent to a *fall* in the covariance component (recall that j is machine's age, which is inversely related to productivity) and a rise in the variance of wages.

[73] This result is related to the intertemporal trade-off intrinsic in the separation decision: choosing to remain on the old vintage improves the current wage (no vintage-specific skill is lost), but worsens future wages because in the next period the worker will have older knowledge, with low degree of transferability. As γ goes up, the expected future wage loss from holding old skills increases faster than the current wage gain, inducing the worker to optimally anticipate its separation decision.

Under this optimal separation rule, the equilibrium level of wage dispersion is given by

$$var(\widetilde{w}) = \gamma^2 \left[\frac{1}{4} + \frac{1}{2}\tau \left(\tau - 1 \right) \right], \tag{27}$$

so it is increasing in γ whenever the variance is well defined (positive). In particular, the equilibrium displays $var(\widetilde{h}) = \frac{\gamma^2\tau^2}{2}$, and $cov(\widetilde{h}, j) = \frac{\gamma\tau}{4}$. The variance of skills is increasing in γ since a higher γ reduces the skill transferability of the bottom-end workers (h_{10}), while not affecting the skill level of the top end workers (h_{01}). The covariance between skills and age of technology is also increasing in γ, a force that restrains inequality because it worsens equilibrium sorting in the economy. The reason is that a larger γ reduces the skills of workers moving to the new technology relatively more than the skills of workers moving to old technologies.

A common criticism of this class of models is that the degree of churning in the labor market (i.e., labor mobility or job reallocation) has to rise in order to generate more volatile earnings, whereas the empirical literature documents no significant rise in labor mobility (Neumark, 2000).[74] This is a misconception. One way to unravel this issue exploits the equivalence between cross-sectional wage dispersion and individual wage instability in a model with ex-ante equal and infinitely lived agents. A technological acceleration has two effects. First, it curtails skill transferability, thereby increasing wage losses upon separation. Second, it reduces the average skill level of workers who find themselves, on average, on the steeper portion of a concave learning curve, which in turn implies higher wage growth on the job. Both these forces tend to raise individual earnings volatility, for any given level of labor mobility. Violante (2002) offers some evidence of wage losses upon separation and wage growth on the job being larger in the 1980s than in the 1970s and shows that a calibrated full-scale version of this model can account up to 90 percent of the rise in wage instability in the U.S. economy, while at the same time implying only a very modest rise in equilibrium labor turnover.

[74]The empirical literature on labor mobility contains partly opposing results: whereas Jovanovic and Rousseau (2004b) find a significant decline of labor mobility since the 1970s, Kambourov and Manovskii (2004) find that occupational mobility increased since the 1970s.

7.3.1 Occupation-Specific Human Capital

Occupation-specific experience may be one of the least transferable components of human capital, and a change in occupational mobility can have a big impact on the wage structure. Kambourov and Manovskii (2004) document an increase in occupational mobility in the United States from 16 percent in the early 1970s to 19 percent in the early 1990s.[75] Based on a calibration exercise Kambourov and Manovskii argue that 90 percent of the rise in residual inequality (i.e., in both the permanent and the stochastic component) is due to increased occupational mobility.

The authors build a model of occupation-specific human-capital accumulation based on the equilibrium search framework of Lucas and Prescott (1974). At any one time workers can work in one occupation only. Workers choose their occupation based on their occupation-specific experience. When working in an occupation, workers increase their specific experience, and they lose some of this experience when moving between occupations. A worker's wage in a given occupation depends on the specific experience and the occupation's productivity.

The productivity of occupations is subject to shocks, and increased variability of these shocks directly increases wage variability. However, the total impact of occupational productivity shocks on wage inequality depends on the occupational choice response of workers. Workers in an occupation whose productivity declines choose to move in search of better occupations, and, by so doing, they dampen the effect of the shock on inequality. When the increased variance of productivity shocks is accompanied by decreased persistence –as conjectured by the authors– workers in occupations hit by moderately negative shocks may choose not to switch occupation because occupations which look profitable today may turn quickly into unproductive ones. This latter effect amplifies the direct effect of the initial shocks.[76]

[75] Kambourov and Manovskii use occupational data from the PSID at the three-digit level, including almost 1000 occupational groups.

[76] The model of Bertola and Ichino (1995), discussed in section 8, generates increased wage inequality through a similar mechanism.

7.3.2 A Precautionary Demand for General Skills

Gottschalk and Moffitt (1994) found that the transitory component of inequality is larger (and increased more) for low-education workers. Gould et al. (2001) model this phenomenon using a vintage human capital model where risk-averse workers choose their level of education. They study an economy where workers are ex-ante heterogeneous with respect to permanent innate ability, and the return to college education is increasing in ability. High-ability workers obtain a college education that provides them with general skills which do not depreciate as technology advances. Low-ability workers do not acquire general skills in college; rather, they acquire technology-specific experience through on-the-job learning. Here, we refer to workers with a college education as skilled and to workers without a college education as unskilled.

Gould et al. (2001) consider a shock to the economy that simultaneously increases the rate of embodied technological change and the ex-ante variance of technological progress across jobs.[77] This shock increases the "precautionary" demand for college education, since holding technology-specific skills becomes more risky. The lowest ability threshold for college graduates falls, and thus permanent inequality increases within skilled workers and falls within the group of unskilled workers. At the same time, the rise in the variance of embodied technological change means that "skill erosion" has a bigger impact on the relative wages of unlucky and lucky unskilled workers, so the increase in their wage variance is mostly determined by transitory components.

This mechanism relies on the assumption that the variance of technical progress is heteroskedastic in the sense that it rises with its mean. We know very little about this property: Cummins and Violante (2002) analyze the whole cross-industry distribution of equipment-embodied technical change for 62 industries in the United States from 1947 to 2000 and find little evidence of changing variance, although the mean grows substantially over the period. However, they document a rise in the cross-sectoral variance of the "technological

[77] This view of the past 30 years as being a period of high "turbulence" is also present in several models of the differential labor market performance between the United States and Europe, see Section 8.

gap" between average capital and leading-edge machines.[78] According to the transferability technology (24), the technological gap closely measures the degree of skill erosion of an average worker displaced in a given industry.

7.3.3 Explaining the Fall in Real Wages

Interestingly, in a set of model economies with vintage human capital (Helpman and Rangel, 1999, Gould et al. 2001, Violante, 2002, or Kambourov and Manovskii, 2004), during the transition to the new steady state, and notwithstanding the technological acceleration, the fall in the average skill level of the workforce can generate a temporary slowdown in average wage growth and a fall in the real value of wages at the bottom of the distribution—two facts that have been documented extensively for the period of interest.

To illustrate this point, let us return to the model from Section 7.3. Note that in an equilibrium where workers separate every period—as assumed—each skill type represents one fourth of all workers. The four skills types are reported in expression (25) . It is immediate to see that the normalized average log level of skills is $-\tau\gamma$, and thus it falls unambiguously when γ increases. This opens the interesting possibility that, in the model, the average wage could decrease along the transition following a technological acceleration.

Suppose that at time t the economy is in steady state with $\gamma = \gamma_L$ (and with the productivity of the new machine normalized to 1). The average log wage is then $\widetilde{w}_t = -\tau\gamma_L - \gamma_L/2$. Suppose now that γ rises to γ_H. Then, some simple algebra shows that in the next period, the average log wage is

$$\widetilde{w}_{t+1} = \frac{\gamma_H}{2} - \frac{\tau}{2}(\gamma_L + \gamma_H) = \widetilde{w}_t - \frac{\tau}{2}(\gamma_H - \gamma_L) + \frac{1}{2}(\gamma_L + \gamma_H).$$

Thus, despite the technological acceleration, the average wage decreases along the transition if $\tau > (\gamma_L + \gamma_H)/(\gamma_H - \gamma_L)$, that is, if τ or the increase in γ are large.

An alternative explanation for the fall in real wages –which does not depend on vintage human capital– is advanced by Manuelli (2000) within a frictional labor market model where workers have bargaining power and can seize a fraction of the firm's future stream of profits,

[78] See section 3.2.1 for a formal definition of the technological gap.

through wage negotiations. Consider what happens when it is announced that: 1) a new technology will be available in the future; but 2) the incumbent firms will be able to adopt it only with some probability (as in Greenwood and Jovanovic, 1999). Existing firms will anticipate a future increase in wages, driven by the new, more productive entrants. Hence, there will be a transitional phase before the arrival of the new technology, where the market value of the incumbent firms will fall and, with them, the wages they currently pay.

7.4 Random Matching vs. Directed Search as Source of Luck

So far, we have analyzed economies where the friction is due to random matching. Wong (2003) argues that models with random matching can have counterfactual implications.[79] It is well known that in a matching model with two types of workers (skilled and unskilled) and two types of firms (high-tech and low-tech), there can be multiple equilibria (Sattinger, 1995). There are equilibria with perfect sorting where skilled (unskilled) workers are matched with high- (low-) tech firms and equilibria that display some degree of "mismatch." In the latter class of equilibria, luck plays a role as skilled workers, ex-ante equal, can end up in jobs with different productivities. Suppose output is the product between efficiency of capital, z_i, where $i = l, h$ and $z_h > z_l$, and efficiency of labor, h_j, where $j = s, u$ and $h_s > h_u$, i.e., $y_{ij} = z_i h_j$. A wave of skill-biased technical change (or a capital-embodied technological acceleration) that increases the relative productivity of high-tech jobs (i.e., the ratio z_h/z_l) makes high-tech firms more picky in their choice of workers, as now the same skill differences translate in larger output differences. The equilibrium with mismatch is less likely to survive. When the economy switches to the equilibrium with perfect sorting, luck-driven inequality among ex-ante equal workers falls to zero.[80]

One of the key reasons why the model has this counterfactual prediction is that, due to random matching, prices (wages) have no signaling value. Shi (2002) analyzes exactly the same framework (a two-worker, two-firm economy) but he replaces Nash bargaining and

[79] See also Albrecht and Vroman (2002) for a similar environment.

[80] The argument in Wong (2003) regarding models with random matching is quite general; e.g., it applies in the model by Acemoglu (1999). From equation (20) of Section 5, note that as h_s rises (skill-biased technical change), the pooling, or mismatch, equilibrium is less likely to survive, so within-group inequality falls.

random matching with wage posting and directed search, following the alternative approach of "competitive search" (Moen, 1997). His conclusion is that random matching is not essential for technical progress to leverage the effect of luck in the labor market: directed search works equally well.

In this environment, skilled workers only apply to high-tech jobs, while unskilled workers apply to both types of jobs. Ex ante, every unskilled worker is indifferent between jobs, but inequality is generated ex post. Since high-tech firms give always priority to skilled applicants, unskilled workers applying for high-tech jobs are less likely to become employed than are unskilled workers applying for low-tech jobs. Therefore unskilled workers applying for high-tech jobs have to be offered higher wages than in low-tech jobs.

With free entry, a rise in the relative productivity of high-tech jobs (skill-biased technical change) induces the creation of more high-tech vacancies. More unskilled workers become attracted to the high-tech sector and in equilibrium their job finding probability in the high-tech sector falls, so wages rise. In the meantime, fewer unskilled workers stay in the low-tech sector, so their wages fall. In sum, wage inequality among ex-ante equal workers rises with the degree of skill bias in technology.

Can one conclude that directed search models are more suitable than random search models for studying problems where heterogeneity is crucial, such as wage inequality? The answer depends on the dimension of inequality studied. Directed search seems a more reasonable assumption when the trait determining heterogeneity is observable (e.g., education, general experience), whereas random matching fits better in the analysis of wage inequality when the source of heterogeneity is not directly observable (e.g., ability or vintage-specific skills).

8 Technology-Policy Complementarity: United States vs. Europe

A large portion of this chapter has been dedicated to the analysis of a number of different economic models designed to decipher the dynamics of the U.S. wage distribution over the

past three decades, in light of changes in technology.

In this section we expand our viewpoint to include other dimensions of labor market inequality, which allows us to contrast the U.S. experience with the European experience. In Section 2 we documented that while wage inequality soared in the United States, both the labor share of income and the unemployment rate remained remarkably stable there. In sharp contrast, in most of the large continental European economies, the wage structure did not change much at all, while the labor share fell substantially and unemployment increased steadily. In particular, the increase in European unemployment largely reflects longer durations rather than higher unemployment incidence.

8.1 The Krugman Hypothesis

Why have we observed such different outcomes for two regions of the world standing at a similar level of development and, therefore, being subject to very similar aggregate shocks? Are we witnessing a sort of *devil's bargain*, i.e., a trade-off between inequalities: low unemployment can only be achieved by paying the price of soaring wage inequality? And, if so, what determines the position of each country along this trade-off?

In Table 2 we report, for the set of countries from Table 1, some indexes of the rigidity of various labor market institutions reproduced from Nickell and Layard (1999). The conclusion is unambiguous: compared to the United States, continental Europe has stricter employment protection legislation, more generous and longer unemployment benefits, less decentralized wage bargaining, and more binding minimum wage law.

TABLE 2

The large majority of papers in the literature have taken the data exhibited in Table 2 as uncontroversial evidence that the reason for the observed differences can be found in the differences in labor market institutions between United States and Continental Europe. Krugman (1994) was probably the first to provide a simple formalized model of this hypothesis. Simply put, the interaction between a severe technological shock and rigid European institutions have induced an adjustment through equilibrium *quantities* of labor (i.e., the

employment distribution), whereas in the flexible U.S. labor market, the adjustment occurred through *prices* (i.e., the wage distribution).

Several authors have tried to test the Krugman hypothesis econometrically. The typical analysis is based on a cross-country panel of institutions and shocks, i.e., it allows for changing institutions over time, beyond aggregate shocks. A statistical model linking shocks and institutions to the dynamics of unemployment and wage inequality is estimated to evaluate the role of shocks and institutions, first separately and then interacted. The shocks considered are usually of technological nature and are measured through changes in measured TFP and changes in the labor share of income, possibly capturing a form of capital-biased technical change. In all cases the shock is assumed to be common across countries.

Blanchard and Wolfers (2000) argue that changing institutions alone have little explanatory power. The performance of the statistical model in explaining cross-country patterns of unemployment rates improves once shocks and institutions are interacted: an equal-size technological shock has differential effects on unemployment when labor market institutions differ. Bertola et al. (2001) provide further evidence for this view. Bentolila and Saint-Paul (1999) also study the evolution of the labor share across OECD countries since 1970. Using panel data techniques, they find that in the presence of institutions that promote wage rigidity, shocks that reduce employment also significantly reduce the labor share of income. One common problem in this empirical literature is that the results are, in general, not robust to the chosen specification.[81]

Another problem of this methodology is that the economic mechanism behind the interaction between technology and policy is not explicit. Consistently with the approach we took in the chapter so far, we will devote more space to quantitative analyses based on "structural" equilibrium models. In the rest of this section, we present the various frameworks the literature has explored to understand the interactions between technological progress and labor market institutions in shaping the various dimensions of inequality. We have grouped

[81] The recent results in Nickell and Nunziata (2002) seem to support an explanation of cross-country unemployment differentials largely based on changing institutions, with a common technological shock playing only a minor role.

these frameworks into six categories, according to the type of technological shock modeled: 1) a rise in microeconomic turbulence, linked to some fundamental change in technology, 2) a slowdown in total factor productivity, 3) an acceleration in the rate of capital-embodied productivity improvements, 4) skill-biased technical change, 5) a technological innovation whose adoption is endogenous, and 6) the structural transformation from manufacturing to services.

8.2 Rise in Microeconomic Turbulence

In Section 2 we have documented that roughly one-half of the rise in cross-sectional wage differentials in the United States is not associated to a higher return to permanent skills. Rather, it is due to increased wage "instability" over the workers' life time. In other words, transitory idiosyncratic shocks to labor productivity and wages have become more important over time (Gottschalk and Moffitt 1994). These larger temporary wage movements constitute important evidence that there has been a rise in the degree of microeconomic turbulence in the U.S. economy.

More evidence comes from the firm side. Campbell et al. (2001) show that the cross-sectional variability of individual stock returns has trended upward from 1962 to 1997. Chaney et al. (2003) and Comin and Mulani (2003) use Compustat firm-level data to demonstrate that the firm-level volatility of real variables, such as investment and sales, has gone up from 1970-1975 to 1990-1995. Overall, these papers provide snapshots, from very different angles, of an economy where idiosyncratic turbulence and volatility have risen to a high level.

Bertola and Ichino (1995) and Ljungqvist and Sargent (1998, 2003) argue that a rise in microeconomic turbulence that interacted with more or less rigid institutions can explain the U.S.-Europe dichotomy. Interestingly, the former authors identify wage rigidity and strict employment protection laws as the culprits, while the latter emphasize the generosity of unemployment benefits. Note, though, that one key premise behind these theories is that the surge in turbulence is common to the United States and Europe. We are not aware of any empirical work documenting trends in microeconomic instability in continental Europe.

Currently, this represents a limit for this class of explanations.

8.2.1 The Role of Wage Rigidity

The framework proposed by Bertola and Ichino (1995) is inspired by the Lucas and Prescott (1974) island-model of equilibrium unemployment. The economy is populated with a measure, L, of risk-neutral workers and a measure one of firms, indexed by $i \in [0,1]$. Each firm is subject to idiosyncratic productivity shocks that follow a two-state Markov chain taking values $\left(A^G, A^B\right)$, with $A^G > A^B$, and with transition probability, p, that the state (good, G, and bad, B) changes. When labor mobility is perfect, employment adjusts across good and bad firms to equalize wage differentials, and a unique market-clearing wage rate arises in equilibrium, i.e., there is no wage inequality.

Consider now the case where wages are flexible, but where workers have to pay a fixed moving cost, $\kappa > 0$, to change firms (this is the U.S.-like economy). In any period, workers observe the productivity level in all firms, but moving takes one period. Hence, when they start working, productivity might change. It is easy to see that the value functions of a worker in good- and bad-state firms, respectively, are

$$W^G = w^G + \frac{1}{1+r}\left[pW^B + (1-p)\,W^G\right], \tag{28}$$

$$W^B = \begin{cases} w^B + \frac{1}{1+r}\left[pW^G + (1-p)\,W^B\right] & \text{if staying,} \\ w^B - \kappa + \frac{1}{1+r}\left[pW^B + (1-p)\,W^G\right] & \text{if moving.} \end{cases} \tag{29}$$

If workers leave bad firms in equilibrium, the marginal worker has to be indifferent between staying in a B firm or moving, yielding

$$W^G - W^B = \frac{1+r}{1-2p}\kappa.$$

Using (28) and (29) together with this condition, one arrives at the expression for equilibrium wage inequality:

$$w^G - w^B = \frac{r+2p}{1-2p}\kappa. \tag{30}$$

On the one hand, the closer p is to 0, the more permanent are productivity changes. This justifies a large amount of wage-equalizing mobility, and hence there is smaller ex-post wage

inequality across firms. On the other hand, the larger is the degree of volatility in the economy (the closer p is to $1/2$), the riskier it is to move for a worker, as the new firm can quickly turn into the B state, and the cost κ is wasted. In this case, mobility will be low and the ex-post wage differential will increase.

Now consider the same experiment in a Europe-like economy where wages are rigid, i.e., where $w^B = w^G = w$, and where firing costs are prohibitively high, so that employment at every firm is constant at $\bar{l}$. To analyze this situation, Bertola and Ichino assume that firm i has a linear marginal revenue product $\pi(l^i) = z^i - \alpha l^i$, so that the marginal values for a firm in the G and B state of a unit of labor, respectively, are

$$V^G = A^G - \alpha\bar{l} - w + \frac{1}{1+r}\left[pV^B + (1-p)V^G\right], \text{ and}$$

$$V^B = A^B - \alpha\bar{l} - w + \frac{1}{1+r}\left[pV^G + (1-p)V^B\right].$$

In an equilibrium with free-entry, the hiring firm in the G state will have $V^G = 0$. Hence, the system above can be easily solved for $\bar{l}$ to give

$$\bar{l} = \frac{A^G - w}{\alpha} - \left(\frac{p}{r+2p}\right)\left(\frac{A^G - A^B}{\alpha}\right), \tag{31}$$

which shows that a rise in p that increases the degree of turbulence in the rigid economy will reduce average employment, i.e., it will increase the unemployment rate, $L - \bar{l}$. The reason is straightforward: when firms are constrained in their ability to shed labor in the face of a negative shock, they will be very cautious in hiring new workers even in the high-productivity state. Note, in fact, that the larger is the productivity differential $A^G - A^B$ across states, the higher will average unemployment in the economy be.

In conclusion, a similar increase in economic uncertainty induces more caution in workers' mobility and larger wage differentials in an economy with flexible wages whereas it leads to more caution in firms' hiring and lower average employment in an economy with rigid wages and costly layoffs. This result remains qualitative, as the authors did not try an exploration of the quantitative importance of their mechanism. In particular, it would be of interest to study by how much labor turnover needs to decline in order to generate a rise in wage

inequality of the magnitude observed in the U.S. economy. Interestingly, as mentioned earlier, Jovanovic and Rousseau (2004b) document a substantial downward trend in labor mobility in the United States, from 50 percent in 1970 to 35 percent in 2000.

8.2.2 The Role of Welfare Benefits

Ljungqvist and Sargent (1998, 2003) propose an alternative mechanism based on the standard search model of unemployment (McCall 1970). Here, we present a stripped-down version of their argument. Consider an unemployed worker with skill level, h, who searches for a job, sampling wage offers every period from the stationary distribution, $F(w)$, with finite support $[\underline{w}, \bar{w}]$. Her skill level, when unemployed, decays at the geometric rate, δ, whereas, when employed, skills remain unchanged. Employment is an absorbing state (no exogenous breakup of jobs), and workers discount the future at rate r. Unemployment benefits are equal to b. The values of employment and unemployment for a worker of skills h are

$$W(w,h) = \frac{wh}{r}, \text{ and}$$

$$U(h) = b + \frac{1}{1+r}\int_{\underline{w}}^{\overline{w}} \max\left\{U(h'), W(w,h')\right\} dF(w)$$

$$s.t. \qquad h' = (1-\delta)h$$

respectively. The value of employment is simply the discounted present value of earnings, wh; the value of unemployment is given by the unemployment benefit plus the discounted future value of search with the lower skill levels $(1-\delta)h$. At the reservation wage, $w^*(h)$, the values of employment and unemployment are equalized, $U(h) = W(w^*(h), h)$. Standard algebra yields the following characterization of the reservation wage,

$$w^*\left(\frac{r+\delta}{1+r}\right) = r\frac{b}{h} + \frac{(1-\delta)}{1+r}\int_{w^*}^{\overline{w}} [1 - F(w)]\, dw. \tag{32}$$

Ljungqvist and Sargent model the increased turbulence in the economy as a rise in the "skill obsolescence" parameter δ. The introduction of a new technological paradigm, or an acceleration in the rate of technological change, can lead to a higher rate of obsolescence, insofar as skills are at least partly technology-specific (recall our discussion in Section 7).

It is straightforward to show, through simple comparative statics, that w^* falls with δ: a worker aware that her skills will become obsolete faster during unemployment chooses optimally to reduce her time spent searching and decreases her reservation wage. As a result, the unemployment duration falls.

However, an increase in δ has an equilibrium effect on the distribution of workers across skills: the average skill level in the population falls, and one can show that the reservation wage declines in the skill level, i.e., $dw^*/dh < 0$. The key behind this result is that the unemployment benefits, b, do not depend on the *current* skill level, h, of the unemployed workers, whereas wage offers are naturally linked to h. A fall in h worsens the value of the average wage offer relative to the value of remaining unemployed with benefits b. Thus, both the reservation wage and unemployment duration increase.[82] The net effect of these two forces is qualitatively ambiguous, and only a quantitative analysis can determine which force is paramount. Note that it is easy to show that the derivative, dw^*/dh, is increasing (in absolute value) in b. Thus, in Europe-like economies with more generous benefits, the second effect tends to be stronger.

Ljungqvist and Sargent embed this simple mechanism in a much richer and detailed model. They calibrate the increase in turbulence to reproduce average earnings losses upon separation of the size estimated in the labor economics literature and show that in economies with generous welfare state (high b), the rise in microeconomic uncertainty brings about a surge in unemployment comparable to the one observed in continental Europe, with all the increase explained by longer durations, as the data suggest. In a "laissez-faire" economy with low b, the faster rate of skill obsolescence barely has any effect.

A related explanation is set forth by Marimon and Zilibotti (1999). In their model, unemployment insurance has the standard result of reducing employment, but it also helps workers find a suitable job. They construct two artificial economies which only differ by the degree of unemployment insurance and assume that they are hit by a common technological shock which enhances the importance of "mismatch". This shock reduces the proportion of

[82] One can easily generalize the model to allow b to depend on past earnings (thus, on *past* skills when employed) and the mechanism described would still be in place. This is what Ljungqvist and Sargent do.

jobs which workers regard as acceptable in the economy with unemployment insurance, and unemployment doubles in the Europe-like economy.

In the Ljungqvist-Sargent and Marimon-Zilibotti frameworks, the shock-policy interaction operates entirely through the *labor supply* side. These authors essentially argue that unemployment in Europe went up because, for the jobless, it was more beneficial to collect unemployment insurance than to work at a low wage, given that technological change made their skills obsolete (or made it difficult to use them on the current jobs).

8.3 Slowdown in Total Factor Productivity

A decline of TFP growth rates, such as measured for the United States and Europe after the mid-1970s (see Section 2) can reduce employment in a matching framework through the standard "capitalization effect." Consider the decision of a firm to create a job: the firm will compare the set-up cost with the discounted present value of profits. In a growing economy, where technical change is disembodied and benefits all firms equally, a productivity slowdown increases the "effective rate" at which profits are discounted and discourages the creation of new jobs (Pissarides 2000).

den Haan et al. (2001) evaluate this explanation quantitatively within the context of a standard matching model, à la Mortensen-Pissarides (1998). They find that for this channel to have a significant effect on unemployment, one needs to put restrictions on the shape of the cross-sectional distribution of firms' productivities. Since useful data to test these restrictions are scant, the mechanism remains largely unexplored.

Interestingly, in the same paper the authors argue that once the Ljungqvist and Sargent mechanism is embedded into a model with endogenous job destruction, the comparative statics for increased turbulence are reversed, i.e., unemployment falls. The reason is that as the speed of skill obsolescence rises, workers become more reluctant to separate, and job destruction falls.[83] This force dominates the effect described in the previous section.

[83]Recall that the original Ljungqvist and Sargent model is a standard search model where separations occur exogenously. Hence, workers are unable to respond to a negative shock hitting their job. In a matching model with wage bargaining, the workers can allow the firm to keep a larger fraction of output in order to avoid a separation in the face of a shock.

Ljungqvist and Sargent (2003) counter-argue that such an economic mechanism would be relevant only if every worker who separates (including those who quit voluntarily) were hit by faster skill obsolescence. In their view, a more reasonable assumption is that only the workers who suffer an exogenous layoff see their skills decreasing, in which case the original result in Ljungqvist and Sargent (1998) remains intact.

8.4 Acceleration in Capital-Embodied Technical Change

Several measures of embodied technical change suggest that the rate of technical change accelerated around the mid-1970s in the U.S. economy (see Section 2, especially Figure 2). A recent OECD study (Colecchia and Schreyer 2002) measures the decline in relative price for several high-tech equipment items across various countries in Europe from 1980 to 2000 and concludes that European countries experienced an acceleration quantitatively comparable to the United States. Jorgenson (2004, Table 3.5) measures the growth in the quality of the aggregate stock of capital across some OECD countries and finds that, even though the United States had the fastest average annual growth (1.5 percent from 1980 to 2001), Germany and Italy were quite close, with 1.3 percent and 1.1 percent annual growth rates, respectively.

Hornstein et al. (2003a) study precisely whether the interaction between an acceleration in capital-embodied growth, common between the United States and Europe, and certain labor market institutions whose strength differs between the United States and Europe, can explain the simultaneous evolution of the three dimensions of labor market inequalities quantitatively: the unemployment rate, the labor share, and wage inequality.

Their environment builds on the matching model with vintage capital developed by Aghion and Howitt (1994).[84] Consider a continuous-time economy populated by a stationary measure one of ex-ante equal, infinitely lived workers who supply one unit of labor inelastically. Workers are risk-neutral and discount the future at rate r. Production requires

[84]For expositional purposes, we simplify the framework in Hornstein et al. (2003a) substantially here. In particular, in the equilibrium of the original model, there are vacant firms with old vintages of machines, while here we make the standard assumption of matching models that all vacant firms embody the leading-edge technology.

one machine and one worker. Machines are characterized by their age, a, translating into match productivity, $e^{-\gamma a}$, where γ is the rate of technological progress embodied in capital.[85]

At any time firms can freely enter the market and post a vacancy at a cost, κ. Then they proceed to search for a worker in a frictional labor market governed by a standard constant-returns-to-scale matching function. Once matched, they produce and share output with the worker in a Nash fashion, with ξ denoting the bargaining power of the worker. At age $\bar{a}$ (determined endogenously), capital is scrapped and the job is destroyed.[86] Two key labor market policies are modeled explicitly: unemployment benefits b, and an employment protection system that combines a hiring subsidy T and an equal-size firing tax upon separation.

As is standard in this framework, it is possible to reduce the equilibrium of the model to two key equations—the job creation condition and the job destruction condition—in two unknowns, θ and $\bar{a}$. These equations, respectively, read

$$\begin{aligned} \kappa &= q(\theta)(1-\xi)S(0;\bar{a}) \\ e^{-\gamma\bar{a}} &= b + p(\theta)\xi S(0;\bar{a}) - (r-\gamma)T. \end{aligned} \tag{33}$$

Here, $q(\theta)$ and $p(\theta)$ are the meeting probabilities for firms and workers, respectively, expressed as a function of the vacancy-unemployment ratio, θ. We denote by $S(0,\bar{a})$ the "surplus" of a match of age 0, conditional on destruction taking place at age $\bar{a}$: the surplus is the value of the relationship for the parties (the discounted present value of the output stream), net of their outside options. Clearly the surplus is increasing in $\bar{a}$, as a longer match yields a bigger surplus.

The job-creation curve states that vacancies are created (and $q(\theta)$ falls) until the expected return of the marginal vacancy equals its cost, κ. The job-destruction curve states that, at age $\bar{a}$, the pair is indifferent between continuing operating the machine, which gives output

[85] As usual, we normalize all variables concerning a vintage a machine relative to the corresponding variable of the newest machine.

[86] Productivity improvements enter the economy only through new capital. This is the typical Schumpeterian "creative-destruction" mechanism, which is at the heart of unemployment in this class of models. As mentioned in section 7.1, Hornstein et al. (2003b) show that if one takes the view that technical progress can also benefit old machines, i.e., if old machines can be "upgraded" into new ones at a cost, then the model yields quantitatively similar results.

$e^{-\gamma\bar{a}}$, and separating, which yields the respective outside options for worker and firm (zero in equilibrium for the firm, because of free entry), net of the firing tax.

FIGURE 6

Figure 6 depicts the comparative statics of a rise in γ in a rigid economy (high b, high T) and in a flexible economy ($b = T = 0$) in the $(\theta, \bar{a})$ space.[87] Note that a low value for $\bar{a}$ corresponds to high separation rate and unemployment incidence, whereas a low value for θ corresponds to long unemployment durations. Thus, the two axes depict the two dimensions of equilibrium unemployment. To illustrate the result more sharply, we have chosen values for b and T in the rigid economy such that the initial equilibrium in the two economies is the same. This is possible since generous benefits and strict employment protection have offsetting effects on job destruction, while they are neutral on job creation, as evident from equations (33). The model is therefore consistent with an initial situation where, originally, the labor markets of the United States and Europe looked alike, as the data for the 1960s show. Figure 6 illustrates that a rise in γ has a dramatically different impact across the two economies, especially regarding the amplitude of the shift in the job destruction curve.

To understand intuitively the economic forces at work, it is useful to think of the acceleration in equipment-embodied technology as an "obsolescence shock." As the rate of productivity growth of new capital accelerates, existing capital-worker matches—which have old vintages of capital—become obsolete faster. In the United States, this loss of economic value is to a higher extent borne by workers, whose wages fall in order to keep firms from scrapping capital and breaking up earlier to invest in better machines.

In Europe, however, labor payments are kept artificially high by generous unemployment benefits and by rents on firing costs, which make wages downwardly rigid. As a result, firms must bear the initial adjustment by destroying matches earlier and creating fewer jobs. The corresponding sharp increase in unemployment greatly improves the relative bargaining position of firms, which can now push workers closer to their outside option, thus reducing

[87] Once we recognize that $p'(\theta) > 0$, $q'(\theta) < 0$, and $S'(\cdot, \bar{a}) > 0$, understanding the slope of the two curves in (33) is immediate.

the labor share of output. Since the outside option is constant across all workers, this force also limits the rise in wage inequality that comes about with faster technical change because of larger productivity differentials across machines.[88] Thus, in response to a technological acceleration, an economy with rigid, European-like institutions would experience a higher unemployment rate, a more pronounced decline in the labor share, and a slower rise in wage inequality than would be observed in a more flexible economy.

Quantitatively, a permanent rise in the rate of capital-embodied productivity growth of the magnitude observed in the data can replicate a large fraction of the differential increase in the unemployment rate and of the capital share between the flexible U.S. economy and the rigid Europe-like economy (with the increase in unemployment taking place along the duration margin, as in the data). Wage inequality increases in the U.S.-like economy and declines in the rigid economy, but the changes generated by the model are rather small (recall our discussion of section 7.1).

8.5 Skill-Biased Technical Change

A number of explanations for the rise in wage inequality in the United States—many of which we have reviewed in Section 3—build on the idea of skill-biased technical change. Could this type of technological advancement, interacted with more rigid institutions, also be at the origin of the rise in European unemployment?

Mortensen and Pissarides (1999) explore this question in a model where the economy is populated by a finite number of types of workers, ex-ante different in their skill (productivity level). Skill is observable (e.g., education), and all workers endowed with the same skill level are segmented in their own labor market, which is modeled as frictional with a standard matching function governing the meeting process.

In this model, unemployed workers receive welfare benefits which are partly proportional to their wage (and skill level), and partly lump-sum. The equilibrium unemployment is decreasing and convex in the skill level: low-skill markets have higher unemployment, as

[88] This mechanism, which is based on technological heterogeneity and the existence of quasi-rents for workers, is the same as that analyzed in Section 7.2.

benefits represent a form of wage rigidity that is more binding at low levels of skills. As benefits become more generous, the convexity becomes more pronounced.

The skill-biased shock is introduced as a mean-preserving spread in the skill distribution, calibrated to match the rise in wage inequality in the economy with low benefits, like the United States. The model predicts a sharp surge in unemployment in the economy with generous benefits, due to the convex equilibrium relationship between unemployment and the skill level. A crucial ingredient of the Mortensen-Pissarides mechanism, which is present also in the Hornstein et al. (2003a) setup, is that welfare benefits are not fully proportional to wages and productivity; rather, they have a "flat", lump-sum component. If they were fully proportional, every skill market would just be a rescaled version of the highest-skill market, with the same unemployment rate. Hansen (1998) studies the institutional details of the welfare state in several European countries and argues that flat "social assistance" benefits are an important component of these welfare systems.[89]

Finally, the Mortensen and Pissarides model has the counterfactual implication that the rise in unemployment is concentrated among the low-skilled workers, whereas Nickell and Bell (1996) and Gottschalk and Smeeding (1997), among others, conclude that data from many European countries support the conclusion that unemployment rose proportionately across the entire skill spectrum.

8.6 Endogenous Technology Adoption

A careful look at Table 1 shows a non-monotonic evolution of the labor share in many European countries: the labor share rose between 1965 and 1980, only to decline sharply afterwards. In some countries this pattern is striking. In Portugal, for example, the labor share skyrocketed from 56 percent to 75 percent in the period 1965-1980, and then plunged to 68 percent by 1995. Blanchard (1997) and Caballero and Hammour (1998) proposed an

[89] Another key assumption is that markets are segregated across skills. This setup allows the model to avoid the criticism by Wong (2003) that skill-biased technical change can reduce the amount of mismatch and decrease within-group inequality. Mortensen and Pissarides describe their workers as differing in educational attainment. Given the observability of education, modeling firms as able to direct their search (and segregate the economy) seems appropriate.

explanation for these dynamics based on the idea that technological advancement responds to the relative cost of factor inputs.

The late 1960s and early 1970s witnessed a rapid evolution of capital-labor relationships in favor of labor in many European countries: "pro-labor" measures were introduced with the objective of consolidating unions' power, increasing the generosity and coverage of unemployment benefits, making economically motivated dismissals harder to justify.[90] The result was, in the language of Caballero and Hammour, an "appropriability shock" that shifted bargaining power away from capital.

In a model where the technological menu for capital-labor substitutability is fixed in the short run, but endogenous in the long run, one will observe an initial rise in the labor share as a result of such a shock. However, as time goes by, more and more firms respond to this institutional "pro-labor" push by introducing new technologies that substitute capital for labor. Therefore, in the long run the capital-labor ratio rises, and both the labor share and employment decline, as observed in the last two decades in Europe.

Why do the U.S. data not display the same pattern? According to Blanchard (1997), since the initial appropriability shock was much smaller, so was the response of capital. A natural question arises, if one follows this logic through: is it only a coincidence that the technological change away from unskilled labor was biased towards capital in Europe and toward skilled labor in the United States?

According to Acemoglu (2003a) the direction of the bias in technological innovations is endogenous (see also our discussion in Section 3) and institutional differences can be key in explaining different biases between the United States and Europe. Consider a flexible economy, like the United States, where firms can either produce with one unit of skilled labor with productivity h_s or one unit of unskilled labor with productivity h_u, where $h_u < h_s$. Output is $y_i = h_i$, with $i = u, s$, and the wage paid to the worker is simply a fraction, ξ, of output. Firms can also choose to pay a fixed cost, κ, and adopt a new technology that

[90]The "French May" in 1968 and the "Hot Italian Autumn" of 1969 are stark manifestations of the power of the labor movements in that period of European history.

increases output by a factor $1 + A < h_s/h_u$. Consider first an equilibrium where

$$\kappa > (1 - \xi) Ah_s > (1 - \xi) Ah_u,$$

so that it is not profitable for firms to implement the new innovation, and wage inequality is simply given by $w_s/w_u = h_s/h_u$. Suppose now that, due to technological progress, the cost of capital decreases to a new value, $\kappa' < \kappa$, such that it is always profitable to adopt it for skilled workers, but not for unskilled workers, i.e.,

$$(1 - \xi) Ah_s > \kappa' > (1 - \xi) Ah_u.$$

As a result of this adoption decision, wage inequality jumps to the higher level $(1 + A) h_s/h_u$ in the U.S. economy.

Consider now an alternative economy, like Europe, where, because of some institutional constraint, wages cannot fall below a fixed level, $\bar{w}$, where $\xi h_s > \bar{w} > \xi (1 + A) h_u$, so that the constraint is binding for the unskilled workers, even in the case of adoption, but never for the skilled workers.

Whenever the new cost level, κ', satisfies

$$Ah_u > \kappa' > (1 - \xi) Ah_u$$

in Europe, the new technology will be adopted also with unskilled workers; this is an effect of the minimum wage constraint. The intuition for this result is that, since firms in Europe pay a fixed wage $\bar{w}$ to the unskilled workers, whether or not they adopt the new technology, the institutional constraint makes the firm the residual claimant on output, once $\bar{w}$ is paid. The new technology increases output without changing the wage payment, and thus it may be optimal to adopt in an economy with wage rigidity and not to adopt in an economy with wage flexibility, with the obvious implication that inequality will not increase in Europe.[91]

Formalized models, where the direction of technical change is endogenous, are still in their infancy: in the case of this application to the U.S.-Europe comparison, one important

[91] This hypothesis is also consistent with the fact that, at least until the impressive productivity surge of 1995-2000, labor productivity grew faster in Europe (e.g., in France, Germany, Italy, and the United Kingdom) compared to the United States (Jorgenson 2004, Table 3.16).

extension would be verifying if this result survives when the "institutional wage rigidity" is endogenized so that it can respond to changes in the technological environment.

8.7 Sectoral Transformation

The standard approach to the U.S.-Europe differentials is built on comparing the diverging dynamics in the *unemployment* rate. Rogerson (2004) argues that the analysis of relative unemployment rates is misleading, and if one focuses instead on *employment-population ratios*, new insights surface.

In particular, Rogerson shows two new features of the data: 1) the relative deterioration of European employment starts as early as in the 1950s, whereas unemployment rates start diverging in the mid-1970s; 2) the deterioration of European unemployment is largely explained by the differential in manufacturing employment growth.[92]

These facts lead Rogerson to focus on the importance of the structural transformation occurring in the economy, i.e., the secular pattern of reallocation of resources across broad sectors of the economy: first from agriculture to manufacturing, and then from manufacturing to services. Expressed in terms of the shocks-institutions paradigm that we have highlighted in this section, the relevant shock is the transformation of modern economies into service-driven economies, and the relevant institutions are those which hampered the full development of a service sector in Europe.

Although this new approach is still in its infancy, and as such it lacks a quantitative assessment within a rigorous equilibrium model, it appears to be quite promising.

8.8 Discussion

Nickell and Layard (1999), in a widely cited piece in the most recent edition of the *Handbook of Labor Economics,* carefully review the empirical literature and conclude that time spent worrying about the effects of several labor market institutions on cross-country unemployment differentials is largely wasted, given these effects seem small and are often even

[92]The concept of "relative deterioration" refers to the difference between the U.S. variable (employment rate or unemployment rate) and its European counterpart.

ambiguous in sign. From the perspective of the research surveyed in this section, however, it seems that when institutional differences are studied in conjunction with technological change, the results are more encouraging.

Of course, much is still far from being well understood. First, once we recognize that the interaction of shocks and institutions is important, what are the key common shocks and the crucial institutional differences that can account for the facts? One would, for example, like to see a unified structural equilibrium framework where several shocks and institutions are jointly analyzed in order to investigate which shock-policy interaction is quantitatively important and which is not.

Second, in answering this question, more "discipline" is needed in the quantitative analysis. Often, the approach in the literature is to calibrate the shock by matching either the rise in wage inequality or the fall in the labor share. We maintain the view that changes in employment/unemployment, wage inequality, and income shares are intimately related and must be explained jointly: they are dimensions along which the model should be evaluated rather than calibrated. Thus, the shock should be calibrated, as much as possible, using independent observations. The use of data on technological change such as that for the relative price of equipment goods is such an example.

Third, it is important to note that we are not aware of any quantitative model of a rigid Europe-like economy that can generate a rise in equilibrium unemployment which is similar across all skill levels, which is what the data suggest.

Fourth, the literature is split between labor-supply models (Ljungqvist and Sargent; Marimon and Zilibotti) and labor-demand models (Bertola and Ichino; Caballero and Hammour; Hornstein et al.). Obviously, interpreting the European and U.S. labor market outcomes in terms of "labor demand" or "labor supply" is not mutually exclusive. In a theoretical framework with elements of vintage human capital and vintage physical capital, an embodied technological acceleration will also worsen the rate of skill obsolescence—exactly as in Ljungqvist and Sargent's paper. The next generation of investigations of the European (un-)employment puzzle should bring together supply and demand forces and allow a joint

evaluation of their respective strength.

9 Welfare and Policy Implications

In traditional growth theory, technological progress is largely associated with productivity advancements, reflected in improvements in average wages, from which it would follow that there are welfare gains. While the first generation of growth models is based on the representative-agent assumption, the model economies we studied in these chapter are built on a heterogenous-agents model. By raising the wage differential between more and less skilled workers (between-group inequality) and by amplifying the amount of labor market uncertainty faced by ex-ante equal households in the economy (residual inequality), in these economies technological change can lead to welfare costs, at least for certain groups of workers, and it has first-order implications for policy. In what follows we give an account of some early work on the subject.

9.1 Lifetime Earnings Inequality

The majority of the empirical investigations on rising inequality in the United States focus on the cross-sectional distribution of wages and earnings. Friedman (1982) argues that data on cross-sectional inequality at a point in time are difficult to interpret, as they provide no information on the degree of economic mobility: the same distribution can be generated either by a "dynamic society" or by a "status society".

A better measure of inequality, which incorporates some of Friedman's concerns, is provided by the distribution of lifetime earnings. A stark example of the pitfalls implicit in making welfare and policy statements simply based on distributions at a point in time is provided by Flinn (2002). Flinn compares Italy and the United States and documents that, although the dispersion in cross-sectional yearly earnings inequality in the United States is several times larger than in Italy, the distribution of lifetime earnings in the United States is more compressed due to larger individual variability of labor income and shorter duration of non-employment experiences. In other words, in Friedman's language, Italy somewhat

surprisingly looks more like a "status society" than does the United States.[93]

Two papers, so far, have studied the change in the distribution of lifetime earnings in the United States in the past three decades through the lenses of a structural model.[94] Heckman et al. (1998) solve a deterministic competitive OLG model with endogenous human capital accumulation to study the implications of the widening educational premium for lifetime-earnings inequality across cohorts.[95] Their model implies that the low-educated cohorts entering in the mid-1980s are those suffering the largest drop in lifetime earnings from skill-biased technical change: roughly 11 percent. At the same time, they calculate a rise in lifetime earnings of 6 percent for the college graduates in the same cohort.

Similarly, Bowlus and Robin (2003) use a search model with risk-neutrality, estimated on matched CPS data from 1977 to 1997, to study how changes in wage and employment dynamics over the past thirty years have affected the evolution of lifetime labor income inequality in the U.S. labor market. They find that the median worker suffered only a small decline in present value lifetime earnings, but that there is large heterogeneity across educational groups with lifetime earnings declining by over 25 percent for high-school graduates and increasing by almost 20 percent for college graduates.

These numbers are over twice as large as those in Heckman et al. (1998). One reason is that Heckman et al. model the acquisition of education and the costs associated with schooling explicitly. A large fraction of the changes in lifetime earnings is attributable to the surge in the returns to education: since education in reality is the outcome of a costly investment choice, the difference in earnings alone likely overstates the true welfare differential between the two groups in the analysis of Bowlus and Robin.

[93] Cohen (1999) performs a similar exercise between the United States and France and finds that, using annual wages, inequality in the United States is 60 percent greater than in France, but based on lifetime earnings, the difference reduces to 15 percent.

[94] See Aaronson (2003) for a measurement of changes in lifetime earnings inequality not based on a structural model.

[95] We have discussed a simple version of this model in Section 4.

9.2 Consumption Inequality

There is a definite gain in moving from studying hourly wages to lifetime labor income, if one wants to make inference on welfare. However, one important limit of the studies above is that they effectively assume complete insurance against those transitory income fluctuations that cancel out in the long run and thus do not affect lifetime income. With imperfect insurance against labor market risk, consumption is not determined only by purely permanent shocks that translate one-for-one into permanent income, but the degree of earnings variability and its persistence become important, too. In this sense, consumption is an even better measure of welfare than lifetime earnings.

The evidence based on *Consumption and Expenditure Survey (CEX)* data suggests that consumption inequality rose slightly during the first half of the 1980s (Cutler and Katz, 1992, and Johnson and Shipp, 1997) and has remained roughly stable thereafter (Krueger and Perri, 2002). Interestingly, Blundell and Preston (1998) document that in Britain, where the increase in wage inequality followed a pattern similar to the United States, the rise in consumption inequality was also strong until the early 1980s, but weaker afterwards. This path of consumption inequality is, at first sight, puzzling, especially since wage inequality keeps increasing in the 1990s, albeit at a slower pace. Three explanations for this puzzle have been provided so far.

Krueger and Perri (2002) developed the first formal model to solve this apparent puzzle. They consider an Arrow-Debreu economy with limited enforcement of contracts (Kocherlakota, 1996). In this economy, the degree of insurance market completeness is endogenous and responds to changes in income risk: as income shocks become larger and more persistent, the value of autarky declines, so agents are willing to enter more often into risk-sharing agreements. The central message of Krueger and Perri is that the rise of labor market inequality led to a development in financial markets—in particular the sharp expansion of consumer credit in the 1990s—and to a larger extent of risk sharing, limiting the rise in consumption inequality in this period.

Heathcote et al. (2003) offer an alternative interpretation for this pattern of rising and

then flattening consumption inequality. Through a statistical decomposition of the rise in wage dispersion into permanent and transitory components, they conclude that the relative importance of the two components changes substantially over the sample period. From the late 1970s to around 1990 the permanent component increases sharply, but in the 1990s it ceases to grow, whereas there is a substantial increase in the variance of transitory shocks. A standard overlapping-generations model with "exogenously" incomplete-markets (Huggett 1996) predicts a trajectory for consumption inequality similar to the data: as the shocks become more transitory, they are easier to insure and tend to have a smaller impact on consumption. The finding that the first phase of the rise in inequality (1980s) had a more permanent nature than the second (1990s) is common to a number of empirical studies (Moffitt and Gottschalk, 1994, for the United States, and Dickens, 2000 and Blundell and Preston, 1998 for the United Kingdom). To our knowledge, there is no attempt to link this pattern of persistence with the nature of technical progress.

The third explanation is provided by Attanasio et al. (2003) who argue that once measurement error in the CEX data is properly taken into account, consumption inequality keeps rising also in the 1990s, and, hence, that there is no puzzle.

9.3 Welfare Implications

Studying consumption inequality is a further improvement toward the understanding of the welfare costs of rising inequality, but a complete welfare analysis cannot abstract from leisure.

One approach that has been taken in the literature makes minimal assumptions regarding the structure of the underlying economic model. Krueger and Perri (2003), in an exercise similar in spirit to that in Attanasio and Davis (1996), estimate a stochastic process directly on consumption and leisure data from the CEX and use standard intertemporal preferences to compute the welfare costs of rising inequality. The computation of welfare losses "under the veil of ignorance," i.e., before the worker finds out whether she will be high- or low-skilled, yields numbers between 1 percent and 2 percent, with a difference in the welfare losses between the 90th percentile (net winners) and the 10th percentile (net losers) of just over 10 percent. To put this number in perspective, the estimate of Bowlus and Robin (2003)

is 50 percent.

This approach is based entirely on revealed preferences, and has the advantage that no restrictive assumptions have to be made on the degree and the nature of market completeness. However, without a structural model (like those of Bowlus and Robin, 2003; and Heckman et al. 1998), strong faith must be placed in the reliability of the consumption and hours data from the CEX. In particular, if there are large transitory measurement errors, then one would overestimate the extent of economic mobility and underestimate the welfare losses coming from the change in the wage structure. Moreover, all that can be assessed through this methodology is the welfare cost of changes in consumption and leisure inequality, without knowing exactly what fraction of these changes are attributable to rising wage inequality rather than, for example, tax reforms or changes in financial and insurance markets that occurred over the same period.

A second approach, developed by Heathcote et al. (2003), builds on three steps: 1) an estimation of the dynamics of permanent and transitory components of individual wages over the period of interest, 2) a calibration of an OLG model with endogenous leisure and consumption choices and incomplete markets, 3) simulation of the model to compute the welfare costs of the changes in wage dynamics. This approach, thus, is fully structural, and, as such, it does not rely heavily on survey data on consumption and hours worked. Rather, welfare calculations are based on the changes in the model-generated consumption and leisure paths due exclusively to observed and well-measured changes in the wage process over the period. At the same time, it incorporates a realistic range of insurance avenues (a saving technology, labor supply, and social security) without going as far as assuming complete markets.

According to the calculations of Heathcote et al., (2003), welfare losses "under the veil of ignorance," although varied by cohort, average 2.5 percent across all cohorts, with a peak of 5 percent for the cohorts entered in the mid-1980s. The low-skill workers suffer a loss of 16 percent, and the high-skill workers enjoy a welfare gain of 13 percent. These numbers fall in between the estimates of Bowlus and Robin (2003) and those of Heckman et al. (1998).

Two main conclusions emerge. First, the welfare consequences of the observed rise in labor market risk are quite different across groups of workers: whereas the high-skill, high-educated workers are the winners, the low-skill, low-educated workers are the losers. Second, the ex-ante welfare loss from the rise in labor market risk in the United States is of the order of 2 percent of lifetime consumption, which is a rather large number.

9.3.1 Insurance and Opportunities in the Welfare Analysis of Wage Inequality

The quantitative studies on the welfare consequences of the recent rise in inequality point to a sizeable welfare loss. But does the absence of full insurance always imply a welfare decrease when risk increases? The answer is no. We have already mentioned the case studied by Krueger and Perri (2002) where, with endogenous market incompleteness, a rise in uncertainty can lead to more risk sharing in society and increase welfare. The same result can arise for different reasons in models where the extent of risk-sharing is limited exogenously (Bewley-Aiyagari economies). Consider, as do Heathcote et al. (2004), an economy populated by a measure one of infinitely-lived agents with preferences

$$U = E_0 \sum_{t=0}^{\infty} \beta^t \left[\frac{c_t^{1-\gamma} - 1}{1-\gamma} - \varphi \frac{h_t^{1+\sigma}}{1+\sigma} \right], \tag{34}$$

where $1/\gamma$ is the elasticity of intertemporal substitution and $1/\sigma$ is the Frisch elasticity of labor supply. Each agent i starts with zero wealth and faces log-normal productivity shocks to the efficiency units of labor ω_{it}. Shocks can be decomposed into two orthogonal components:

$$\ln \omega_{it} = \alpha_i + \varepsilon_{it}, \quad \text{with } \alpha_i \sim N\left(-\frac{v_\alpha}{2}, v_\alpha\right), \text{ and } \varepsilon_{it} \sim N\left(-\frac{v_\varepsilon}{2}, v_\varepsilon\right)$$

where α_i is the permanent-uninsurable component and ε_{it} is the transitory-insurable component. Note that the means have been normalized so that a rise in the variance of either component does not affect the average level of efficiency units.

After computing the allocations and substituting them into preferences (34), one can calculate the welfare gain of an increase in the two components of wage uncertainty—expressed

as the equivalent consumption variation. The main finding is that one can obtain an (approximate) closed-form expression for the welfare gain $\mathcal{W}$

$$\mathcal{W} \;=\; \frac{1}{\sigma}\Delta v_{\varepsilon} - \frac{(\gamma-1)+\gamma\,(1+\sigma)}{\gamma+\sigma}\Delta v_{\alpha}. \tag{35}$$

This expression only depends on two elasticity parameters (γ, σ) and on the change in the two variances $(\Delta v_{\alpha}, \Delta v_{\varepsilon})$. The key feature to note, in the above equation, is that the welfare gain is not always negative. For example, as $\gamma \to 0$ (risk-neutrality), the welfare gain is positive and proportional to the rise in overall inequality $(\Delta v_{\varepsilon} + \Delta v_{\alpha})$ through the Frisch elasticity.[96]

To understand this result, one has to keep in mind that there are two distinct effects of a rise in labor market uncertainty. On the one hand, "decreased insurance" induces a welfare loss. For example, as risk-aversion rises with γ or as the permanent-uninsurable component v_{α} expands, the second term becomes larger and the overall welfare gain tends to become negative. On the other hand, "improved production opportunities" induce a welfare gain. In presence of elastic labor supply (σ low), households supply more hours when they face a good productivity shock and enjoy leisure at times of low-productivity. When the variance of productivity shocks rises, this intertemporal behavior can improve households' welfare. The net effect depends on the parameterization of preferences and on the empirical assessment of what fraction of the rise in inequality is insurable.

9.3.2 Discussion

Economists have just started to tackle these issues, and many questions still lie ahead. One key area to explore is the role of the family in determining the welfare implications of the rise in wage inequality. Two offsetting forces are at work. First, there is positive assortative matching between spouses along the skill/education dimension. Second, shocks are imperfectly correlated between spouses (Hyslop 2001). While the first feature amplifies the surge in inequality and worsens welfare inequalities across families, the second establishes

[96] This qualitative result can be reproduced also starting from Cobb-Douglas preferences, albeit the expression in (35) is different.

a role for intra-family insurance in dampening the rise in labor market risk. Only a careful quantitative analysis can determine which force is dominant.

Finally, the current welfare studies abstract from some first-order "social" consequences of the rise in inequality and the fall in the wages of the unskilled, such as the decline in labor market participation for low-educated males (Murphy and Topel 1997), the rise in the crime rate (Kelly 2000), and the decline in the marriage rate (Gould and Paserman 2003).[97]

9.4 Brief Directions for Policy

Welfare losses originating from the rise in U.S. inequality in the past three decades are almost one hundred times larger than the standard estimates of the costs of business cycles (Lucas 2003). In this sense, policies that act by reallocating risk across agents (like social insurance policies) are a macroeconomic priority compared to policies that reduce the impact of aggregate risk (like monetary or fiscal stabilization policies). But among the myriads of possible government interventions, what are the right redistributive policies?

In Sections 3.2 and 7 we discussed two complementary views of the link between technology and inequality. The first of these views is that technological progress in the past three decades has been complementary to certain permanent individual characteristics, such as ability or education (technology-skill complementarity). The second view is that labor market history is scattered with shocks and stochastic events related to the luck of individuals, firms or industries that determine the degree of fanning out of the skill and earnings distributions among ex-ante equal workers. The rapid diffusion of a new technology amplifies the importance of these stochastic factors, increasing overall earnings instability (technology-luck complementarity).

The emphasis we placed on these two approaches is not just for classification purposes, since they have profoundly different policy implications. Insofar as we are interested in designing policies that reduce inequalities among households, models of technology-ability complementarity suggest that the intervention should be targeted early in the life of an

[97] Gould and Paserman argue that the higher male inequality in the United States increased the option value for single women to search longer for a husband.

individual, possibly during childhood when the key components of learning ability are being formed. Models of technology-luck complementarity seem to call for interventions that allow the disadvantaged (or unlucky) workers to rebuild their skill level after a shock, such as displacement due to skill obsolescence, has hit.

Examples of both types of policies are abundant in the U.S. economy.[98] In general, the most recent evaluations of programs entailing expenditures and treatment at early childhood report remarkable success. In contrast, the available evidence indicates that welfare-to-work and training programs directed toward adult workers are rather inefficient, as they generate only modest increases in permanent earnings levels (LaLonde et al. 1999).

According to Heckman (2000), the reason for the divergence in outcomes across these two classes of policies is twofold. First, investments in human capital at old ages are less efficient, since the elderly worker has less time to recoup the investment; second, "learning begets learning," so human capital, skills, and abilities acquired at a young ages facilitate future learning.

In this sense, policymakers should have a life-cycle perspective: lifting the unskilled, displaced adults into skilled status is much easier and more efficient if the same workers have been developing their learning ability throughout childhood and youth, possibly with the help of government intervention. For the more mature low-skilled workers with limited learning ability who are subject to unavoidable wage losses due to biased technological change, targeted wage subsidies can be more effective than retraining programs.

10 Concluding Remarks

This chapter argues that labor market inequalities are shaped by technological change through a variety of economic mechanisms. Within the technology-labor market nexus, however, which of the specific mechanisms we evaluate are most likely to survive the test of

[98] Programs like the Perry Pre-School program and the Syracuse Pre-School program provide intense family development support to disadvantaged children at very young ages (from birth to 5 years). The Harlem program ensures frequent individual teacher-child sessions for children of age 3-5. Several programs for adult retraining of displaced workers were initiated throughout the United States under the Job Training Partnership Act of 1982 and the Economic Dislocation and Worker Adjustment Assistance Act of 1988.

time?

Before answering this question, it is useful to put things in perspective and recall that most of the statements we have made in this chapter are not all meant to represent general insights; rather, they allude to a particular historical episode. Specifically, technology has not always been skill-biased in the past: the transformation from artisanal workplaces to the factory in the 19th century had much the opposite effect (Goldin and Katz, 1998). Moreover, not all the drastic productivity advancements in the past were embodied in equipment: electricity was to a large extent embodied in new structures, as the electrification of production required a whole new blueprint for the plant (Atkeson and Kehoe, 2002). Even in reference to this particular historical episode, there are serious dissenting views on the overall impact of IT on the macroeconomy (e.g., Gordon 2000) and on the role of technology in explaining the observed changes in the U.S. wage structure (e.g., Card and DiNardo, 2002).

In returning to the original question, we identify three rather general categories that we find particularly interesting and plausible.

The first idea is *factor-specificity* of the recent technological advancements. In particular, the embodiment of productivity improvements in equipment capital goods, and the skill-bias of such productivity improvements. Whether in the Nelson-Phelps version of skills as a vehicle of adoption and innovation, or in the version of skills and capital as complementary in production, the skill-bias of the IT revolution is one of the most robust and pervasive in the literature. Skill-biased technical change and capital-skill complementarity are crucial to explain the climb of the skill premium, notwithstanding the continuous growth in the relative supply of skilled labor. A growing and promising avenue of research is on the endogenous determinants of the factor-bias in technological advancements (Acemoglu 2002b, 2003b).

The second idea is *vintage human capital*. The technological specificity of knowledge appears to be an important idea to explain some of the most puzzling aspects of the data such as the rise in within-group or "residual" inequality, the fall of the real wages at the bottom of the skill distribution, the growth in the returns to experience, and the slowdown of output growth in the aftermath of a technological revolution.

The third idea is the *interaction between technology and the organization of labor markets.* Radical technological developments, like those we have witnessed in the past three decades, are bound to interact deeply with the various aspects of the structure of labor markets, like the organization of production within the firm, labor unions, and labor market policies. Through this interaction, the literature has successfully interpreted the move from the Tayloristic to the flatter multi-tasking organizational design of firms, the decline of unionization, and the upward trend in unemployment rate in Europe. In particular, the comparison of the U.S. and European experiences seems a fruitful way of studying this channel.

These ideas are the building blocks of the most successful and influential papers in the first generation of models that we have surveyed in this chapter. Where will the literature go next? We argued in various parts of the chapter that one major weakness of this literature is the scarcity of rigorous quantitative evaluations of the theories proposed. Most of the papers reviewed are qualitative in nature. This is not too surprising, given the young vintage of the literature (which developed only starting from the mid 1990s), and given that, in any field, it naturally takes a long time before a handful of theoretical frameworks emerge as successful and begin to be used for a systematic quantitative accounting of the facts (e.g., the search and matching model in the theory of unemployment, and the neoclassical and the endogenous growth model in the theory of cross-country income differences). In this chapter we have highlighted some features that seem important for a successful theory of the link between technological change and labor market outcomes. Quantitative theory should be a priority within this field of research over the years to come.

References

[1] Aaronson, S. (2003), "The Rise in Lifetime Earnings Inequality Among Men," mimeo, Federal Reserve Board.

[2] Abowd J.M., F. Kramarz, and D.N. Margolis (1999), "High Wage Workers and High Wage Firms," *Econometrica* 67, 251-334.

[3] Abraham A. (2003), "Wage Inequality and Education Policy with Skill-biased Technological Change in OG Setting ," mimeo, Duke University.

[4] Acemoglu, D. (1998), "Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality," *Quarterly Journal of Economics* 113, 1055-1090.

[5] Acemoglu, D. (1999), "Changes in Unemployment and Wage Inequality: An Alternative Theory and Some Evidence," *American Economic Review* 89, 1259-78.

[6] Acemoglu, D. (2002a), "Technical Change, Inequality and the Labor Market," *Journal of Economic Literature* 40, 7-72.

[7] Acemoglu, D. (2002b), "Directed Technical Change," *Review of Economic Studies* 69, 781-810.

[8] Acemoglu, D. (2003a), "Cross-Country Inequality Trends," *Economic Journal* 113, F121-149.

[9] Acemoglu, D. (2003b), "Labor- and Capital-Augmenting Technical Change," *Journal of the European Economic Association* 1, 1-37.

[10] Acemoglu, D. (2003c), "Patterns of Skill Premia," *Review of Economic Studies* 70, 199-230.

[11] Acemoglu, D., and J.-S. Pischke (1999), "The Structure of Wages and Investment in General Training," *Journal of Political Economy* 107, 539-72.

[12] Acemoglu, D., P. Aghion and G.L. Violante (2001), "Deunionization, Technical Change, and Inequality", *Carnegie-Rochester Conference Series on Public Policy* 55, 29-64.

[13] Addison, J.T., R.W Bailey and W.S. Siebert (2004), "The Impact of Deunionization on Earnings Dispersion Revisited", Departmental Working Paper 14172, Department of Commerce, University of Birmingham.

[14] Aghion, P. (2002), "Schumpeterian Growth Theory and the Dynamics of Income Inequality," *Econometrica* 70, 855-82.

[15] Aghion, P. and P. Howitt (1994), "Growth and Unemployment," *Review of Economic Studies* 61, 477-494.

[16] Aghion, P. and P. Howitt (1998), Endogenous Growth Theory, (MIT Press: Cambridge and London).

[17] Aghion, P., P. Howitt and G.L. Violante (2002), "General Purpose Technology and Within-Group Wage Inequality," *Journal of Economic Growth* 7, 315-345.

[18] Albrecht, J. and S. Vroman (2002), "A Matching Model with Endogenous Skill Requirements," *International Economic Review* 43, 283-305.

[19] Atkeson, A. and P.J. Kehoe (2002), "The Transition to a New Economy Following the Second Industrial Revolution," NBER Working Papers 8676

[20] Attanasio, O., E. Battistin and H. Ichimura (2003), "What Really Happened to Consumption Inequality in the US?" mimeo, Institute of Fiscal Studies.

[21] Attanasio, O. and S.J. Davis (1996), "Relative Wage Movements and the Distribution of Consumption," *Journal of Political Economy* 104, 1227-62.

[22] Autor, D., L. Katz and A. Krueger (1998), "Computing Inequality: Have Computers Changed the Labor Market?" *Quarterly Journal of Economics* 113, 1169-1213.

[23] Bahk, B.H. and M. Gort (1993), "Decomposing Learning by Doing in New Plants," *Journal of Political Economy* 101, 561-583.

[24] Baker, G.P. and T.N. Hubbard (2003), "Contractibility and Asset Ownership: On-Board Computers and Governance in U.S. Trucking," *American Economic Review* 93, 1328-1353.

[25] Bartel A.P. and F.R. Lichtenberg (1987), "The Comparative Advantage of Educated Workers in Implementing New Technology," *Review of Economics and Statistics* 69, 1-11.

[26] Bartel, A. and N. Sicherman (1998), "Technological Change and the Skill Acquisition of Young Workers," *Journal of Labor Economics* 16, 718-755.

[27] Basu, S., J. Fernald, N. Oulton and S. Srinivasan (2003), "The Case of the Missing Productivity Growth: Or, Does information technology explain why productivity accelerated in the United States and not in the United Kingdom?," *NBER Macroeconomics Annual* 18 (MIT Press: Cambridge and London, 9-63.

[28] Beaudry, P. and D.A. Green (2003), "Wages and Employment in the United States and Germany: What Explains the Differences?" *American Economic Review* 93, 573-602.

[29] Becker, G. (1973), "A Theory of Marriage", *Journal of Political Economy* 81, 813-846.

[30] Bentolila, S. and G. Saint-Paul (1999), "Explaining Movements in the Labor Share," *Contributions to Macroeconomics* 3 (1).

[31] Bertola, G. and A. Ichino (1995), "Wage Inequality and Unemployment: United States vs. Europe," in B. Bernanke and J. Rotemberg, eds., *NBER Macroeconomics Annual* 10 (MIT Press: Cambridge and London), 13-54.

[32] Bertola, G., F.D. Blau and L.M. Kahn (1997), "Swimming Upstream: Trends in the Gender Wage Differential in 1980s," *Journal of Labor Economics* 15, 1-42.

[33] Bertola, G., F.D. Blau and L.M. Kahn (2001), "Comparative Analysis of Labor Market Outcomes: Lessons for the United States from International Long-Run Evidence," in: A.B. Krueger and R.M. Solow, eds., The roaring nineties: Can full employment be sustained? (Century Foundation Press: New York), 159-218.

[34] Bils, M. and P. Klenow (2004), "Measuring Quality Change and Externalities," in: P. Aghion and S. Durlauf, eds., Handbook of Economic Growth, (North-Holland: Amsterdam).

[35] Blanchard, O. (1997), "The Medium Run," *Brookings Papers of Economic Activity (Macroeconomics)* 2, 89-141.

[36] Blanchard, O. and J. Wolfers (2000), "The Role of Shocks and Institutions in the Rise of European Unemployment: The Aggregate Evidence," *Economic Journal* 110, C1-33.

[37] Bleaney, M. (1996), "Central Bank Independence, Bargaining Structure, and Macroeconomic Performance in the OECD Countries," *Oxford Economic Papers* 48, 20-38.

[38] Blundell, R. and I. Preston (1998), "Consumption Inequality and Income Uncertainty," *Quarterly Journal of Economics* 113, 603-640.

[39] Bolton, P. and M. Dewatripont (1994), "The Firm as a Communication Network," *Quarterly Journal of Economics* 109, 809-839.

[40] Booth, A. (1995), The Economics of the Trade Union, (Cambridge University Press: Cambridge, UK).

[41] Borghans, L., and B. ter Weel (2003), "The Diffusion of Computers and the Distribution of Wages," mimeo, Maastricht University

[42] Bowlus, A.J. and J.-M. Robin (2003), "Twenty Years of Rising Inequality in U.S. Lifetime Labor Income Values," forthcoming *Review of Economic Studies*.

[43] Braverman, L. (1974), Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century. (Monthly Review Press: New York).

[44] Bresnahan, T.F. and M. Trajtenberg (1995), "General Purpose Technologies: 'Engines of Growth'?" *Journal of Econometrics* 65, 83-108.

[45] Breshnahan, T.F., E. Brynjolfsson and L.M. Hitt (2002), "Information Technology, Workplace Organization and the Demand for Skilled Labor: Firm-Level Evidence," *Quarterly Journal of Economics* 117, 339-376.

[46] Brynjolfsson, E. and L.M. Hitt (2000), "Beyond Computation: Information Technology, Organizational Tranformation and Business Performance," *Journal of Economic Perspectives* 14, 23-48.

[47] Caballero, R.J. and M.L. Hammour (1998), "Jobless Growth: Appropriability, Factor Substitution, and Unemployment," *Carnegie-Rochester Conference Series On Public Policy* 48, 51-94.

[48] Cain, L. and D. Paterson (1986), "Biased Technical Change, Scale, and Factor Substitution in American Industry, 1850-1919," *Journal of Economic History* 46, 153-64.

[49] Campbell, J., M. Lettau, B. Malkiel and Y. Xu (2001), "Have Individual Stocks Become More Volatile? An Empirical Exploration of Idiosyncratic Risk," *Journal of Finance* 56, 1–43.

[50] Card, D. (1996), "The Effects of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica* 64, 957-979.

[51] Card, D. (2001), "The Effects of Unions on Wage Inequality in the U.S. Labor Market," *Industrial and Labor Relations Review* 54, 296-315.

[52] Card, D. and J. DiNardo (2002), "Skill Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles," NBER working paper 8769.

[53] Card, D. and T. Lemieux (2001), "Can Falling Supply Explain the Rising Return to College for Younger Men? A Cohort-Based Analysis," *Quarterly Journal of Economics* 116, 705-46.

[54] Caroli, E. and J. Van Reenen (2001), "Skill-Biased Organizational Change? Evidence From a Panel of British and French Establishments," *Quarterly Journal of Economics* 116, 1449-1492.

[55] Caselli, F. (1999), "Technological Revolutions," *American Economic Review* 89, 78-102.

[56] Chaney, T., X. Gabaix and T. Philippon (2003), "The Evolution of Microeconomic and Macroeconomic Volatility," mimeo, NYU Stern.

[57] Chun (2003), "Information Technology and the Demand for Educated Workers: Disentangling the Impacts of Adoption versus Use," *Review of Economics and Statistics* 85, 1-8.

[58] Coase (1937), "The Nature of the Firm," *Economica* 4, 386-405

[59] Cohen, D. (1999), "Welfare Differentials Across French and U.S. Labor Markets: A General Equilibrium Interpretation," CEPREMAP working paper 9904.

[60] Cohen-Pirani, D., and R. Castro (2004), "Why Has Skilled Employment Become so Procyclical after 1985?," mimeo Carnegie-Mellon University.

[61] Colecchia, A. and P. Schreyer (2002), "ICT Investment and Economic Growth in the 1990s: Is the United States a Unique Case? A Comparative Study of Nine OECD Countries," *Review of Economic Dynamics* 5, 408-442.

[62] Comin, D. and Mulani (2003), "Diverging trends in macro and micro volatility: Facts," mimeo NYU.

[63] Cooley, T.F. and E.C. Prescott (1995), "Economic Growth and Business Cycles," in: T.F. Cooley, ed., Frontiers of Business Cycle Research (Princeton University Press: Princeton), 1-38.

[64] Cozzi G., and G. Impullitti (2004), "Technology Policy and Wage Inequality", mimeo NYU

[65] Cummins, J. and G.L. Violante (2002), "Investment-Specific Technological Change in the U.S. (1947-2000): Measurement and Macroeconomic Consequences," *Review of Economic Dynamics* 5, 243-284.

[66] Cutler, D.M. and L.M. Katz (1992), "Rising Inequality? Changes in the Distribution of Income and Consumption in the 1980's," *American Economic Review* 82, 546-51.

[67] David, P.A. (1990), "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox," *American Economic Review* 80, 355-61.

[68] den Haan W.J. (2003), "Temporary Shocks and Unavoidable Transitions to a High-Unemployment Regime," mimeo London Business School.

[69] den Haan, W., C. Haefke and G. Ramey (2001), "Shocks and Institutions in a Job Matching Model", NBER working paper 8463.

[70] Dickens, R. (2000), "The Evolution of Individual Male Earnings in Great Britain: 1975-1995," *Economic Journal* 110, 27-49.

[71] Dickens, W.T., and J.S. Leonard (1985), "Accounting for the Decline in Union Membership: 1950-1980," *Industrial and Labor Relations Review* 38, 323-34.

[72] DiNardo, J., N.M. Fortin and T. Lemieux (1996), "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica* 64, 1001-1044.

[73] Dinopolous and Segerstrom (1999), "A Schumpeterian Model of Protection and Relative Wages, " *American Economic Review* 89, 450-472.

[74] Dooley and Gottschalk (1984), "Earnings Inequality among Males in the United States: Trends and the Effect of Labor Force Growth," *Journal of Political Economy* 92, 59-89.

[75] Duranton, G. (2004), "The Economics of Production Systems: Segmentation and Skill-Biased Change," *European Economic Review* 48, 307-36.

[76] Eckstein, Z. and E. Nagypal (2004), "U.S. Earnings and Employment Dynamics 1961-2002: Facts and Interpretations," mimeo Northwestern University.

[77] Farber, H. and A.B. Krueger (1992), "Union Membership in the United States: The Decline Continues," NBER working paper 4216.

[78] Farber, H.S. and B. Western (2000), "Round Up the Usual Suspects: The Decline of Unions in the Private Sector, 1973-1998," Princeton University, Industrial Relations Sections, working paper 437.

[79] Farber, H.S. and B. Western (2002), "Ronald Reagan and the Politics of Declining Union Organization," *British Journal of Industrial Relations* 40, 385-401.

[80] Flinn, C. (2002), "Labour Market Structure and Inequality: A Comparison of Italy and the U.S." *Review of Economic Studies* 69, 611-45.

[81] Flug, and Hercowitz (2000), "Equipment Investment and the Relative Demand for Skilled Labor: International Evidence," *Review of Economic Dynamics* 3, 461-485.

[82] Freeman, R.B (1988), "Contraction and Expansion: The Divergence of Private Sector and Public Sector Unionism in the United States," *Journal of Economic Perspectives* 2, 63-88.

[83] Freeman, R.B. and E.P. Lazear (1995), "An Economic Analysis of Works Councils," in: J. Rogers and W. Streeck, eds., Works Councils: Consultation, Representation, and Cooperation in Industrial Relations, National Bureau of Economic Research Comparative Labor Markets Series, (University of Chicago Press: Chicago and London), 27-50.

[84] Freeman, R.B. and J.L. Medoff (1984), "What Unions Do: Evidence, Interpretation, and Directions for Research," Harvard Institute of Economic Research Discussion Paper: 1096.

[85] Friedman, M. (1982), Capitalism and Freedom, (Chicago University Press: Chicago).

[86] Galor, O., and O. Moav (2000), "Ability Biased Technological Transition, Wage Inequality Within and Across Groups, and Economic Growth," *Quarterly Journal of Economics* 115, 469-97.

[87] Galor, O., and D. Tsiddon (1997), "Technological Progress, Mobility, and Economic Growth," *American Economic Review* 87, 362-382.

[88] Garicano, L. and E. Rossi-Hansberg (2003), "Organization and Inequality in a Knowledge Economy," mimeo Stanford.

[89] Gittleman, Maury, and Mary Joyce (1996); "Earnings Mobility and Long-Run Inequality: An Analysis Using Matched CPS Data," *Industrial Relations* 35, 180-197.

[90] Goldin, C. and L.M. Katz (1998), "The Origins of Technology-Skill Complementarity," *Quarterly Journal of Economics* 113, 693-732.

[91] Goldin, C. and L.M. Katz (1999), "The Returns to Skill in the United States Across the Twentieth Century," NBER working paper 7126.

[92] Goldin, C. and R. Margo (1992), "The Great Compression: The Wage Structure in the United States at Mid-Century," *Quarterly Journal of Economics* 107, 1-34.

[93] Gordon, R.J. (1990), The Measurement of Durable Good Prices, (University of Chicago Press: Chicago).

[94] Gordon, R.J. (2000), "Does the 'New Economy' Measure Up to the Great Inventions of the Past?" *Journal of Economic Perspectives* 14, 49-74.

[95] Gosling, A. and S. Machin (1995), "Trade Unions and the Dispersion of Earnings in British Establishments, 1980-90," *Oxford Bulletin of Economics and Statistics* 57, 167-84.

[96] Gottschalk, P. and R. Moffitt (1994), "The Growth of Earnings Instability in the U.S. Labor Market," *Brookings Papers of Economic Activity* 2, 217-272.

[97] Gottschalk, P. and T.M. Smeeding (1997), "Cross-National Comparisons of Earnings and Income Inequality," *Journal of Economic Literature* 35, 633-687.

[98] Gould, E.D., O. Moav and B.A. Weinberg (2001), "Precautionary Demand for Education, Inequality and Technological Progress," *Journal of Economic Growth* 6, 285-315.

[99] Gould, E.D. and M.. Paserman (2003), "Waiting for Mr. Right: Rising Inequality and Declining Marriage Rates," *Journal of Urban Economics* 53, 257-281.

[100] Greenwood, J. and B. Jovanovic (1999), "The IT Revolution and the Stock Market," *American Economic Review* 89, 116-122.

[101] Greenwood, J. and A. Seshadri (2004), "Technological Progress and Economic Transformation," in: P. Aghion and S. Durlauf, eds., Handbook of Economic Growth, (North-Holland: Amsterdam).

[102] Greenwood J. and M. Yorukoglu (1997), "1974," *Carnegie-Rochester Conference Series on Public Policy* 46, 49-96.

[103] Greenwood, J., Z. Hercowitz and P. Krusell (1997), "Long-Run Implications of Investment-Specific Technological Change," *American Economic Review* 87, 342-362.

[104] Greenwood, J., A. Seshadri and M. Yorukoglu (2004), "Engines of Liberation," *Review of Economic Studies*, forthcoming.

[105] Griliches, Z. (1969), "Capital-Skill Complementarity," *Review of Economics and Statistics* 5, 465-68.

[106] Haider, S. (2001), "Earnings Instability and Earnings Inequality of Males in the United States: 1967-1991," *Journal of Labor Economics* 19, 799-836.

[107] Hall, R.E. (1973), "The Specification of Technology with Several Kinds of Output," *Journal of Political Economy* 81, 878-892.

[108] Hall, R.E. (2001), "The Stock Market and Capital Accumulation," *American Economic Review* 95, 1185-1202.

[109] Hamermesh, D.S. (1993), Labor Demand, (Princeton University Press: Princeton, NJ)..

[110] Hansen, H. (1998), "Transition from Unemployment Benefits to Social Assistance in Seven European OECD Countries," *Empirical Economics* 23, 5-30.

[111] Hayek, F.A. (1945), "The Use of Knowledge in Society," *American Economic Review* 35, 519-530.

[112] Heathcote, J., K. Storesletten and G.L. Violante (2003), "The Macroeconomic Implications of Rising Wage Inequality in the U.S.," mimeo NYU.

[113] Heathcote, J., K. Storesletten and G.L. Violante (2004), "Insurance and Opportunites: The Welfare Analysis of Wage Dispersion," mimeo NYU.

[114] Heckman, J.J. (2000), "Policies to Foster Human Capital," *Research in Economics* 54, 3-56.

[115] Heckman, J.J., L. Lochner and C. Taber (1998), "Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents," *Review of Economic Dynamics* 1, 1-58.

[116] Helpman, E. and A. Rangel (1999), "Adjusting to a New Technology: Experience and Training," *Journal of Economic Growth* 4, 359-383.

[117] Ho, M.S. and D.W. Jorgenson (1999), "The Quality of the U.S. Workforce," mimeo Harvard University.

[118] Hobjin, B. (2000), "Identifying Sources of Growth," mimeo, Federal Reserve Bank of New York.

[119] Holmes, T. and M.F. Mitchell (2004), "A Theory of Factor Allocation and Plant Size," NBER Working Papers 10079.

[120] Hornstein, A. (1999), "Growth Accounting with Technological Revolutions," *Federal Reserve Bank of Richmond Economic Quarterly* 85 (3), 1-24.

[121] Hornstein, A. (2004), "(Un)Balanced Growth," *Federal Reserve Bank of Richmond Economic Quarterly* 90 *Fall*, 25-45.

[122] Hornstein, A. and P. Krusell (1996), "Can Technology Improvements Cause Productivity Slowdowns?", *NBER Macroeconomics Annual* 11 (MIT Press: Cambridge MA), 209-259.

[123] Hornstein, A. and P. Krusell (2000), "The IT Revolution: Is it Evident in the Productivity Numbers?" *Federal Reserve Bank of Richmond Economic Quarterly* 86 *Fall*, 49-78.

[124] Hornstein, A. and P. Krusell (2003), "Implications of the Capital-Embodiment Revolution for Directed R&D and Wage Inequality," *Federal Reserve Bank of Richmond Economic Quarterly* 89 *Fall*, 25-50.

[125] Hornstein, A., P. Krusell and G.L. Violante (2003a), "Vintage Capital in Frictional Labor Markets", mimeo NYU.

[126] Hornstein, A., P. Krusell and G.L. Violante (2003b), "A Quantitative Study of the Replacement Problem in Frictional Economies," mimeo NYU.

[127] Hoxby C.M. and B.T. Long (1998), "Explaining Rising Income and Wage Inequality Among the College-Educated," mimeo harvard University.

[128] Huggett, M. (1996), "Wealth Distribution in Life-Cycle Economies," *Journal of Monetary Economics* 38, 469-94.

[129] Huggett, M. and S. Ospina (2001), "Does Productivity Growth Fall After the Adoption of New Technology?" *Journal of Monetary Economics* 48, 173-95.

[130] Hulten, C.R. (1992), "Growth Accounting When Technical Change is Embodied in Capital," *American Economic Review* 82, 964-80.

[131] Hyslop, D. (2001), "Rising U.S. Earnings Inequality and Family Labor Supply: The Covariance Structure of Intrafamily Earnings," *American Economic Review* 91, 755-77.

[132] Ingram B. and G. Neumann (1999), "An Analysis of the Evolution of the Skill Premium, " mimeo University of Iowa.

[133] Irwin, D.A. and P.J. Klenow (1994), "Learning-by-Doing Spillovers in the Semiconductor Industry," *Journal of Political Economy* 102, 1200-1227.

[134] Iversen, T. (1998), "Wage Bargaining, Central Bank Independence and the Real Effects of Money", *International Organization*.

[135] Johnson, D. and S. Shipp (1997), "Trends in Inequality Using Consumption-Expenditures in the U.S. from 1960 to 1993," *Review of Income and Wealth* 43, 133-52.

[136] Jones C. (2004), "Growth and Ideas" in: P. Aghion and S. Durlauf, eds., Handbook of Economic Growth, (North-Holland: Amsterdam).

[137] Jorgenson, D.W. (2001), "Information Technology and the U.S. Economy," *American Economic Review* 91, 1-32.

[138] Jorgenson, D.W. (2004), "Accounting for Growth in the Information Age," in: P. Aghion and S. Durlauf, eds., Handbook of Economic Growth (North-Holland: Amsterdam).

[139] Jorgenson, D.W. and K.J. Stiroh (2000), "U.S. Economic Growth at the Industry Level," *American Economic Review* 90, 161-67.

[140] Jorgenson, D.W., F. Gollop and B. Fraumeni (1987), Productivity and U.S. Economic Growth (Harvard University Press: Cambridge MA).

[141] Jovanovic, B. (1998), "Vintage Capital and Inequality," *Review of Economic Dynamics* 1, 497 - 530.

[142] Jovanovic, B. and Y. Nyarko (1995), "A Bayesian Learning Model Fitted to a Variety of Empirical Learning Curves," *Brookings Papers on Economic Activity (Microeconomics)*, 247-99.

[143] Jovanovic, B. and P. Rousseau (2004a), "General Purpose Technologies," in: P. Aghion and S. Durlauf, eds., Handbook of Economic Growth, (North-Holland: Amsterdam).

[144] Jovanovic, B. and P. Rousseau (2004b), "Specific Capital and the Division of Rents," mimeo NYU.

[145] Juhn, C. (1992), "Decline of Male Labor Market Participation: The Role of Declining Market Opportunities", *Quarterly Journal of Economics* 107, 79-121.

[146] Juhn, C., K. Murphy and B. Pierce (1993), "Wage Inequality and the Rise in Returns to Skill," *Journal of Political Economy* 101, 410-442.

[147] Kambourov, G. and I. Manovskii (2004), "Occupational Mobility and Wage Inequality", mimeo University of Pennsylvania.

[148] Katz, L. and D. Autor (1999), "Changes in the Wage Structure and Earnings Inequality," in: O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics*, vol. 3 (North-Holland: Amsterdam), 1463-1555.

[149] Katz, L. and K. Murphy (1992), "Changes in Relative Wages, 1963-1987: Supply and Demand Factors," *Quarterly Journal of Economics* 107, 35-78.

[150] Kelly, M. (2000), "Inequality and Crime," *The Review of Economics and Statistics* 82, 530-539.

[151] Kiley, M.T. (1999), "The Supply of Skilled Labour and Skill-Biased Technological Progress," *The Economic Journal* 109, 708-24.

[152] King, I. and L. Welling (1995), "Search, unemployment, and growth," *Journal of Monetary Economics* 3, 499-507

[153] Kocherlakota, N. (1996), "Implications of Efficient Risk Sharing without Commitment," *Review of Economic Studies* 63, 595-609.

[154] Kremer, M. and E.S. Maskin (1996), "Wage Inequality and Segregation by Skill," NBER Working Papers 5718.

[155] Krueger A.B. and L.H. Summers (1988), "Efficiency Wages and the Inter-Industry Wage Structure," *Econometrica* 56, 259-93.

[156] Krueger, D., and Kumar (2004); "Skill-specific rather than General Education: A Reason for US-Europe Growth Differences?," *Journal of Economic Growth* 9, 167-207.

[157] Krueger, D. and F. Perri (2002), "Does Income Inequality Lead to Consumption Inequality? Evidence and Theory", mimeo Stanford.

[158] Krueger, D. and F. Perri (2003), "On the Welfare Consequences of the Increase in Inequality in the United States," in: M. Gertler and K. Rogoff, eds., *NBER Macroeconomics Annual* 18, (MIT Press: Cambridge, MA).

[159] Krugman, P. (1994), "Past and Prospective Causes of High Unemployment," *Economic Review* (Federal Reserve Bank of Kansas City) 79 (4), 23-43.

[160] Krusell, P., L. Ohanian, J.-V. Rios-Rull and G.L. Violante (2000), "Capital Skill Complementarity and Inequality: A Macroeconomic Analysis," *Econometrica* 68, 1029-1053.

[161] LaLonde, R.J., J.J. Heckman and J. Smith (1999), "The Economics and Econometrics of Active Labor Market Programs," in: O. Ashenfelter and D. Card, eds., Handbook of Labor Economics, Vol 3A (North-Holland: Amsterdam), 1865-2097.

[162] Lee, D. and K. Wolpin (2004), "Intersectoral Labor Mobility and the Growth of the Service Sector", mimeo NYU.

[163] Levy, P.A. (1985), "The Unidimensional Perspective of Reagan's Labor Board", *Rutgers Law Journal* 16, 269-390.

[164] Levy, F. and R. J. Murnane (1992), "U.S. Earnings Levels and Earnings Inequality: A Review of Recent Trends and Proposed Explanations", *Journal of Economic Literature* 30, 1333-1381.

[165] Lillard, L.A. and H.W. Tan (1986), "Training: Who Gets It and What Are Its Effects on Employment and Earnings?" (Santa Monica: RAND Corporation Report R-3331-DOL/RC).

[166] Lindbeck, A. and D.J. Snower (1996), "Reorganization of Firms and Labor Market Inequality," *American Economic Review, P&P* 86, 315-321.

[167] Lindquist, M.J. (2002), "Capital-Skill Complementarity and Inequality in Sweden," mimeo Uiversity of Stockholm.

[168] Lindquist, M.J. (2004), "Capital-Skill Complementarity and Inequality Over the Business Cycle," *Review of Economic Dynamics* 7, 519-540.

[169] Lipsey, R.G., C. Bekar and K. Carlaw (1998), "The Consequences of Changes in GPTs," in: E. Helpman, ed., General Purpose Technologies and Economic Growth (MIT Press: Cambridge MA), 193-218.

[170] Ljungqvist, L. and T.J. Sargent (1998), "The European Unemployment Dilemma," *Journal of Political Economy* 106, 514-50.

[171] Ljungqvist, L. and T.J. Sargent (2003), "European Unemployment and Turbulence Revisited in a Matching Model," mimeo NYU.

[172] Lloyd-Ellis, H. (1999), "Endogenous Technological Change and Wage Inequality," *American Economic Review* 89, 47-77.

[173] Lucas, R.E. Jr. (1993), "Making a Miracle," *Econometrica* 61, 251-72.

[174] Lucas, R.E. Jr. (2003), "Macroeconomic Priorities," *American Economic Review* 93, 1-14.

[175] Lucas, R.E. Jr. and E.C. Prescott (1974), "Equilibrium Search and Unemployment," *Journal of Economic Theory* 7, 188-209.

[176] Machin, S. (1996), "Wage Inequality in the UK," *Oxford Review of Economic Policy* 12, 47-64.

[177] Machin, S. (2000), "Union Decline in Britain," *British Journal of Industrial Relations* 38, 631-45.

[178] Machin, S. (2003), "New Workplaces, New Workers: Trade Union Decline and the New Economy", forthcoming in H. Gospel and S. Wood, eds., The Future of Unions ,Volume 1 (Routledge: London).

[179] Manuelli, R. (2000), "Technological Change, the Labor Market, and the Stock Market", NBER Working Papers 8022.

[180] Marimon, R. and F. Zilibotti (1999), "Unemployment vs. Mismatch of Talents: Reconsidering Unemployment Benefits," *Economic Journal* 109, 266-91.

[181] McCall, J.J. (1970), " Economics of Information and Job Search," *Quarterly Journal of Economics* 84, 113-26.

[182] McConnell, S. (1996), "The Role of Computers in Reshaping the Workforce," *Monthly Labour Review* 119 (August), 3-5.

[183] McGrattan, E., and E.C. Prescott (2003), "Taxes, Regulations, and the Value of U.S. Corporations: A General Equilibrium Analysis," Staff Report 309, Federal Reserve Bank of Minneapolis.

[184] Meghir, C. and L. Pistaferri (2004), "Income Variance Dynamics and Heterogeneity", *Econometrica* 72, 1-32.

[185] Milgrom, P. and J. Roberts (1990), "The Economics of Modern Manufacturing: Technology, Strategy,and Organization," *American Economic Review* 80, 511-528.

[186] Mincer, J. and Y. Higuchi (1991), "Wage Structures and Labor Turnover in the United States and Japan," *Journal of the Japanese and International Economies* 2, 97-133.

[187] Mitchell, M.F. (2001), "Specialization and the Skill Premium in the 20th Century," Staff Report 290, Federal Reserve Bank of Minneapolis

[188] Möbius, M. (2000), "The Evolution of Work", mimeo Harvard.

[189] Moen, E.R. (1997), "Competitive Search Equilibrium," *Journal of Political Economy* 105, 385- 411.

[190] Mortensen, D.T. and C.A. Pissarides (1998), "Technological Progress, Job Creation, and Job Destruction," *Review of Economic Dynamics* 1, 733-53.

[191] Mortensen, D.T. and C.A. Pissarides (1999), "Unemployment Responses to 'Skill-Biased' Technology Shocks: The Role of Labour Market Policy," *Economic Journal* 109, 242-65.

[192] Murphy, K. and R. Topel (1997), "Unemployment and Nonemployment," *American Economic Review P&P* 87, 295-300.

[193] Murphy, K. and F. Welch (1992), "The Structure of Wages," *Quarterly Journal of Economics* 107, 285-326.

[194] Nelson, R.R. and E.S. Phelps (1966), "Investment in Humans, Technological Diffusion, and Economic Growth," *American Economic Review* 56, 69-75.

[195] Neumark, D. (2000), "Changes in Job Stability and Job Security: A Collective Effort to Untangle, Reconcile and Interpret the Evidence," in: D. Neumark, ed., On the Job: Is Long-Term Employment a Thing of the Past? (Russell Sage Foundation: New York), 1-27.

[196] Nickell, S. and B. Bell (1996), "Changes in the Distribution of Wages and Unemployment in OECD Countries," *American Economic Review P&P* 86, 302-08.

[197] Nickell, S. and R. Layard (1999), "Labor Market Institutions and Economic Performance," in: O. Ashenfelter and D. Card, eds., Handbook of Labor Economics, vol. 3C (North Holland: Amsterdam), 3029-84.

[198] Nickell, S. and L. Nunziata (2002), "Unemployment in the OECD since the 1960s: What Do We Know?" mimeo, Bank of England.

[199] OECD Employment Outlook (1996), OECD: Paris.

[200] Oliner, S.D. and D.E. Sichel (2000), "The Resurgence of Growth in the Late 1990s: Is Information Technology the Story?" *Journal of Economic Perspectives* 14, 3-22.

[201] Ortigueira, S. (2002), "The Rise and Fall of Centralized Wage Bargaining," mimeo Cornell.

[202] Piketty, T. and E. Saez (2003), "Income Inequality in the United States, 1913-1998," *Quarterly Journal of Economics* 118, 1-39.

[203] Piore, M.J. and C. F. Sabel.(1984), The Second Industrial Divide, (Basic Books: New York).

[204] Pissarides, C.A. (2000), Equilibrium Unemployment Theory, (MIT Press: Cambridge MA).

[205] Rajan, R. and J. Wulf (2003), "The Flattening Firm: Evidence from Panel Data on the Changing Nature of Corporate Hierarchies," mimeo Chicago GSB.

[206] Robbins D. (1996), "Evidence on Trade and Wages in Developing World," OECD Technical Papers 119.

[207] Rogerson, R. (2004), "Two Views on the Deterioration of European Labor Market Outcomes," *Journal of the European Economic Association* 2, 447-455.

[208] Rosen, S. (1981), "The Economics of Superstars," *American Economic Review* 71, 845-58.

[209] Ruiz-Arranz, M. (2002). "Wage Inequality in the U.S.: Capital-Skill Complementarity vs. Skill-Biased Technological Change," mimeo, Harvard University.

[210] Saint-Paul, G. (2001), "On the Distribution of Income and Worker Assignment under Intrafirm Spillovers, with an Application to Ideas and Networks," *Journal of Political Economy* 109, 1-37.

[211] Sakellaris, P. and D.J. Wilson (2004), "Quantifying Embodied Technical Change", *Review of Economic Dynamics* 7, 1-26.

[212] Sattinger, M. (1995), "Search and the Efficient Assignment of Workers to Jobs," *International Economic Review* 36, 283-302.

[213] Shi, S. (2002), "A Directed Search Model of Inequality with Heterogeneous Skills and Skill-Based Technology," *Review of Economic Studies* 69, 467-91.

[214] Solow, R. (1957), "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics* 39, 312-320.

[215] Solow, R. (1960), "Investment and Technological Progress," in: K. Arrow, S. Karlin, and P. Suppes, eds., Mathematical Methods in the Social Sciences (Stanford University Press: Stanford CA), 89-104.

[216] Thesmar, D. and M. Thoenig (2000), "Creative Destruction and Firm Organization Choice", *Quarterly Journal of Economics* 115, 1201-37.

[217] Violante, G.L. (2002), "Technological Acceleration, Skill Transferability and the Rise in Residual Inequality," *Quarterly Journal of Economics* 117, 297-338.

[218] Weinberg, B.A. (2003a), "Computer Use and the Demand for Women Workers," *Industrial and Labor Relations Review* 53, 290-308.

[219] Weinberg, B.A. (2003b), "Experience and Technology Adoption," Working Paper.

[220] Wong, L.Y. (2003), "Can the Mortensen-Pissarides Model with Productivity Changes Explain U.S. Wage Inequality?" *Journal of Labor Economics* 21, 70-105.

Cross-country labor market data (1965-1995)

		1965	1970	1975	1980	1985	1990	1995	Change
	Unemp. Rate	0.018	0.011	0.017	0.029	0.045	0.054	0.061	0.043
Austria	Labor share	0.698	0.679	0.717	0.694	0.665	0.646	0.645	-0.053
	Inequality				0.820	0.790	0.870	0.880	0.060
	Unemp. Rate	0.023	0.022	0.064	0.114	0.111	0.110	0.142	0.120
Belgium	Labor share		0.667	0.729	0.730	0.682	0.685	0.676	0.009
	Inequality					0.660	0.650	0.640	-0.020
	Unemp. Rate	0.014	0.016	0.061	0.093	0.085	0.112	0.103	0.089
Denmark	Labor share	0.736	0.723	0.732	0.706	0.677	0.635	0.605	-0.131
	Inequality				0.760	0.770	0.770		0.010
	Unemp. Rate	0.025	0.021	0.050	0.051	0.047	0.121	0.167	0.142
Finland	Labor share	0.738	0.711	0.762	0.730	0.723	0.733	0.680	-0.058
	Inequality				0.890	0.920	0.940	0.930	0.040
	Unemp. Rate	0.020	0.027	0.049	0.079	0.101	0.105	0.115	0.095
France	Labor share	0.688	0.674	0.707	0.710	0.645	0.618	0.603	-0.085
	Inequality				1.210	1.210	1.240	1.230	0.020
	Unemp. Rate	0.010	0.011	0.037	0.060	0.075	0.078	0.099	0.089
Germany	Labor share	0.685	0.703	0.703	0.704	0.667	0.658	0.637	-0.048
	Inequality				0.870	0.830	0.830	0.810	-0.060
	Unemp. Rate	0.047	0.055	0.078	0.112	0.164	0.146	0.120	0.073
Ireland	Labor share	0.828	0.842	0.835	0.833	0.763	0.715	0.645	-0.183
	Inequality								
	Unemp. Rate	0.041	0.043	0.051	0.070	0.099	0.096	0.120	0.079
Italy	Labor share	0.669	0.687	0.711	0.690	0.656	0.653	0.606	-0.063
	Inequality				0.850	0.830	0.770	0.970	0.120
	Unemp. Rate	0.010	0.018	0.038	0.080	0.081	0.062	0.071	0.061
Netherlands	Labor share	0.656	0.687	0.705	0.661	0.623	0.619	0.624	-0.032
	Inequality					0.920	0.960	0.950	0.030
	Unemp. Rate	0.016	0.015	0.018	0.026	0.030	0.056	0.049	0.034
Norway	Labor share	0.750	0.771	0.782	0.757	0.739	0.713		-0.037
	Inequality				0.720	0.720	0.680		-0.040
	Unemp. Rate	0.040	0.024	0.065	0.079	0.070	0.051	0.073	0.033
Portugal	Labor share	0.562	0.615	0.873	0.751	0.673	0.679	0.680	0.118
	Inequality								
	Unemp. Rate	0.028	0.030	0.059	0.161	0.200	0.196	0.230	0.202
Spain	Labor share	0.763	0.780	0.788	0.756	0.679	0.669	0.616	-0.147
	Inequality								
	Unemp. Rate	0.018	0.022	0.019	0.028	0.021	0.052	0.079	0.061
Sweden	Labor share	0.724	0.716	0.745	0.711	0.691	0.693	0.630	-0.095
	Inequality				0.750	0.760	0.730	0.790	0.040
	Unemp. Rate	0.019	0.025	0.044	0.089	0.091	0.086	0.079	0.060
UK	Labor share	0.693	0.699	0.698	0.694	0.690	0.712	0.692	-0.002
	Inequality				0.920	1.050	1.150	1.200	0.280
	Unemp. Rate	0.040	0.058	0.076	0.099	0.089	0.103	0.096	0.056
Canada	Labor share	0.716	0.660	0.652	0.634	0.630	0.666	0.659	-0.057
	Inequality				1.240	1.390	1.380	1.330	0.090
	Unemp. Rate	0.038	0.054	0.070	0.083	0.062	0.066	0.055	0.017
USA	Labor share	0.685	0.695	0.675	0.678	0.665	0.666	0.670	-0.015
	Inequality				1.180	1.350	1.380	1.470	0.290
Europe	Unemp. Rate	0.024	0.024	0.047	0.076	0.087	0.095	0.110	0.086
Average	Labor share	0.708	0.712	0.753	0.726	0.683	0.670	0.637	-0.062
	Inequality				0.859	0.841	0.844	0.900	0.040

Note: Data on unemployment rates are from Blanchard and Wolfers (2000). Data on labor shares are from Blanchard and Wolfers (2000) except the 1995 entry for Austria, Denmark, Ireland and Portugal which was computed directly from OECD data. Inequality is measured as the 90-10 log-wage differential for male workers. The data are taken from the OECD Employment Outlook (1996, Table 3.1). Austria: the measure is the 80-10 differential and data in the 1985 column are for 1987. Belgium: the measure is the 80-10 differential and data in the 1995 column are for 1993. Denmark: 1985 and 1990 columns are for 1983 and 1991 respectively. Finland: data in the 1985 column are for 1986. Germany: data in the 1985 and 1995 columns are for 1983 and 1993 respectively. Italy: data in the 1985, 1990 and 1995 columns are for 1984, 1991 and 1993 respectively. Netherlands: the measure of inequality is for males and females. Norway: data in the 1985 and 1990 columns are for 1983 and 1991 respectively. Moreover, the measure of inequality is for males and females. Portugal: data in the 1990 and 1995 columns are for 1989 and 1993 respectively. Canada: data in the 1980 and 1985 columns are for 1981 and 1986 respectively. For all countries, except US and UK, data in the 1995 column are for 1994. Europe average: unweighted mean of European countries, except UK.

Table 1: Data on the evolution of the labor share, the unemployment rate, and wage inequality across OECD countries from 1965-1995.

Cross-country institutions data (1984-1995)

	Labor Standards	Employment Protection	Union Density	Bargaining Centralization	Ratio of min. to avg. wage	Benefit Repl. Rate	Benefit Duration
Austria	5	16	46.2	17	0.62	0.50	2.0
Belgium	4	17	51.2	10	0.60	0.60	4.0
Denmark	2	5	71.4	14	0.54	0.90	2.5
Finland	5	10	72.0	13	0.52	0.63	2.0
France	6	14	9.8	7	0.50	0.57	3.0
Germany	6	15	32.9	12	0.55	0.63	4.0
Ireland	4	12	49.7	6	0.55	0.37	4.0
Italy	7	20	38.8	5	0.71	0.20	0.5
Netherlands	5	9	25.5	11	0.55	0.70	2.0
Norway	5	11	56.0	16	0.64	0.65	1.5
Portugal	4	18	31.8	7	0.45	0.65	0.8
Spain	7	19	11.0	7	0.32	0.70	3.5
Sweden	7	13	82.5	15	0.52	0.80	1.2
UK	0	7	39.1	6	0.40	0.38	4.0
Canada	2	3	35.8	1	0.35	0.59	1.0
USA	0	1	15.6	2	0.39	0.50	0.5
Europe Average	5.15	13.77	44.52	10.77	0.54	0.61	2.38

Note: Data are taken from Nickell and Layard (1999), Tables 6, 7, 9, 10. Labor standards are summarized in an index whose max value is 10 and refers to labor market standards enforced by legislation. The employment protection index ranges from 1 to 10. Union density is measured as a percentage of all salary earners. Centralization is an index where 17 corresponds to the most centralized regime. Benefit duration is in years. Europe average: unweighted mean of European countries, except UK.

Table 2: Data on various labor market institutions across OECD countries. Averages for the period 1985-1995.

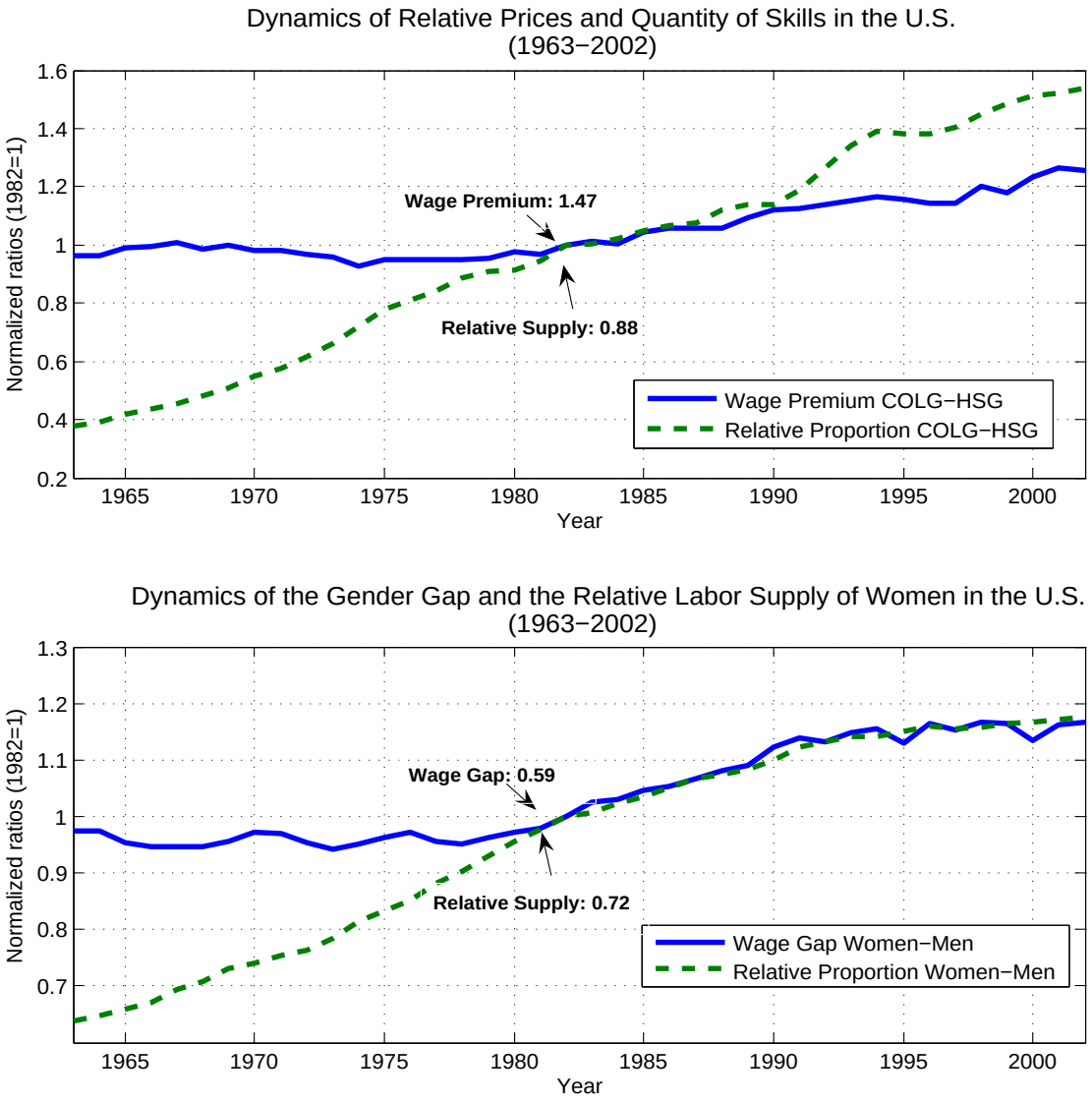


Figure 1: The top panel depicts the evolution of the skill premium (average wage of college graduates relative to the wage of high-school graduates) and of the relative quantity of skilled workers, from 1963-2002. The bottom panel depicts the evolution of the gender gap (average wage of female workers relative to the wage of male workers) and of the relative quantity of female workers, over the same period of time.

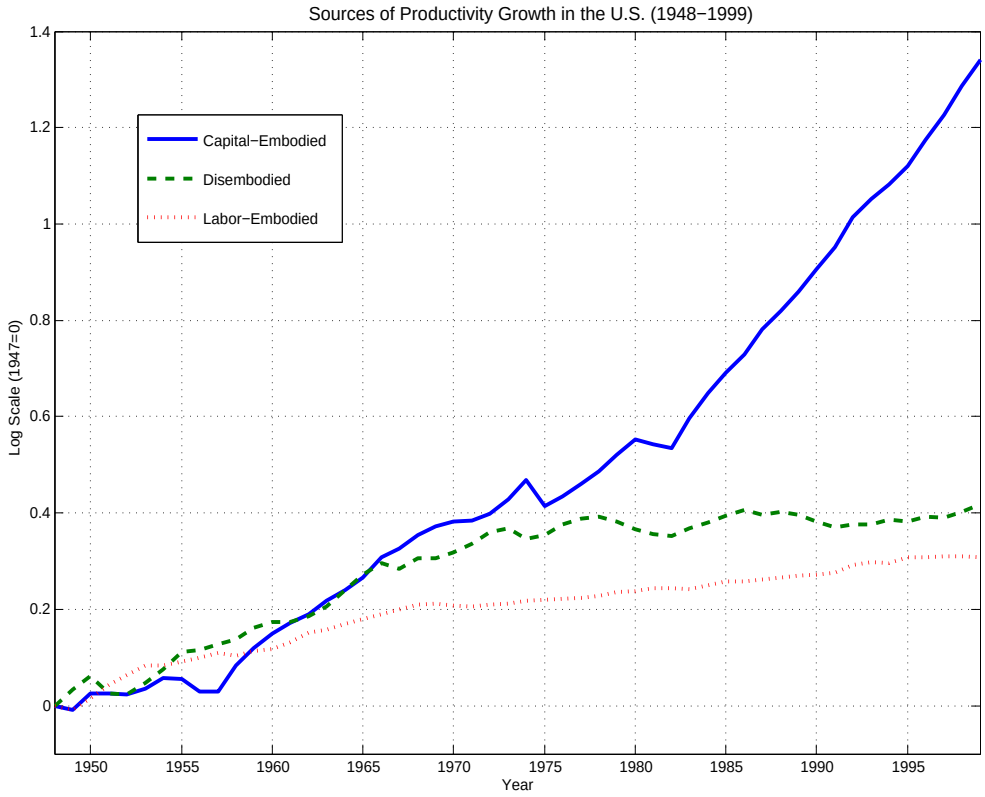


Figure 2: The figure depicts the dynamics of three sources of productivity growth in the post-war U.S. economy: disembodied, capital-embodied, and labor-embodied. Source: Cummins and Violante (2002).

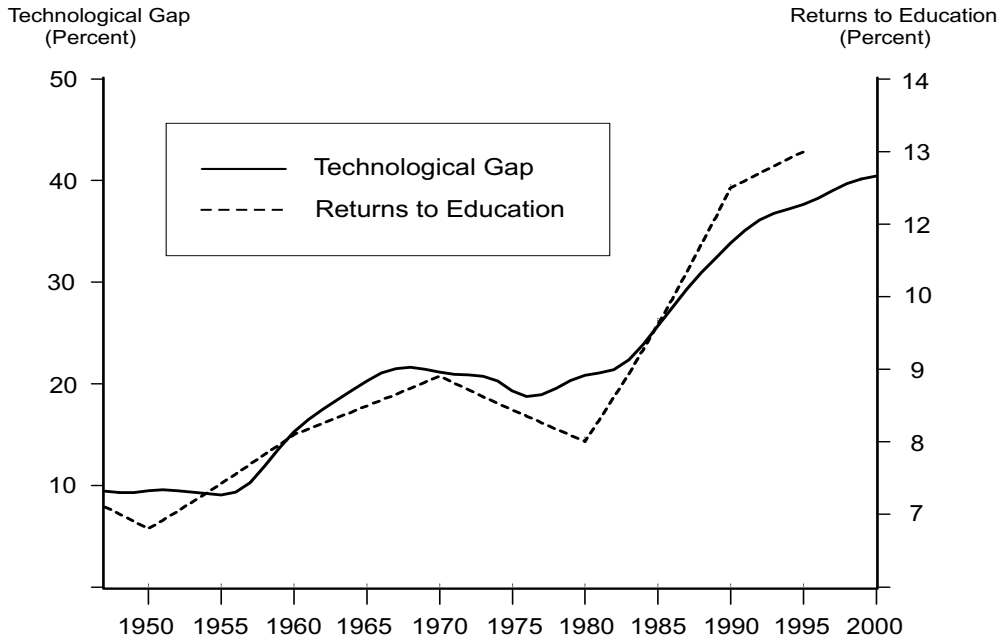


Figure 3: The figure illustrates the joint dynamics of the returns to education and the technological gap (1947-2000) in the U.S. economy. The figure is reproduced from Cummins and Violante (2002).

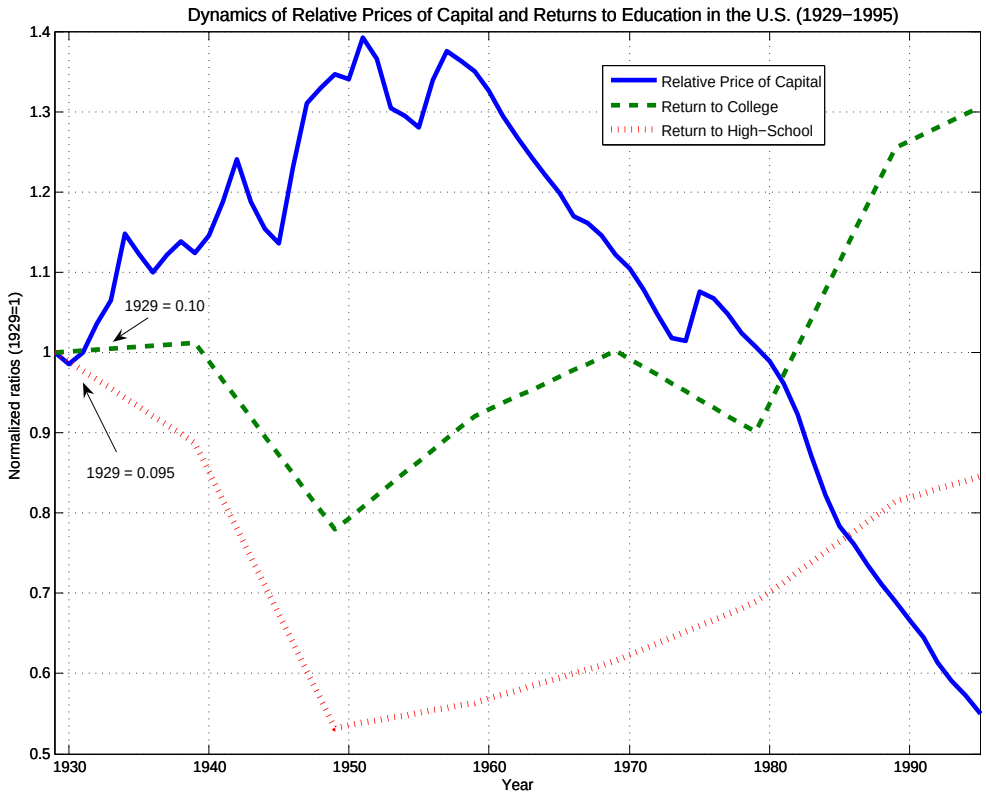


Figure 4: The figure depicts the dynamics of the relative price of capital and the returns to education from 1929-1995 in the U.S. economy. Source: Cummins and Violante (2002) and Goldin and Katz (1999).

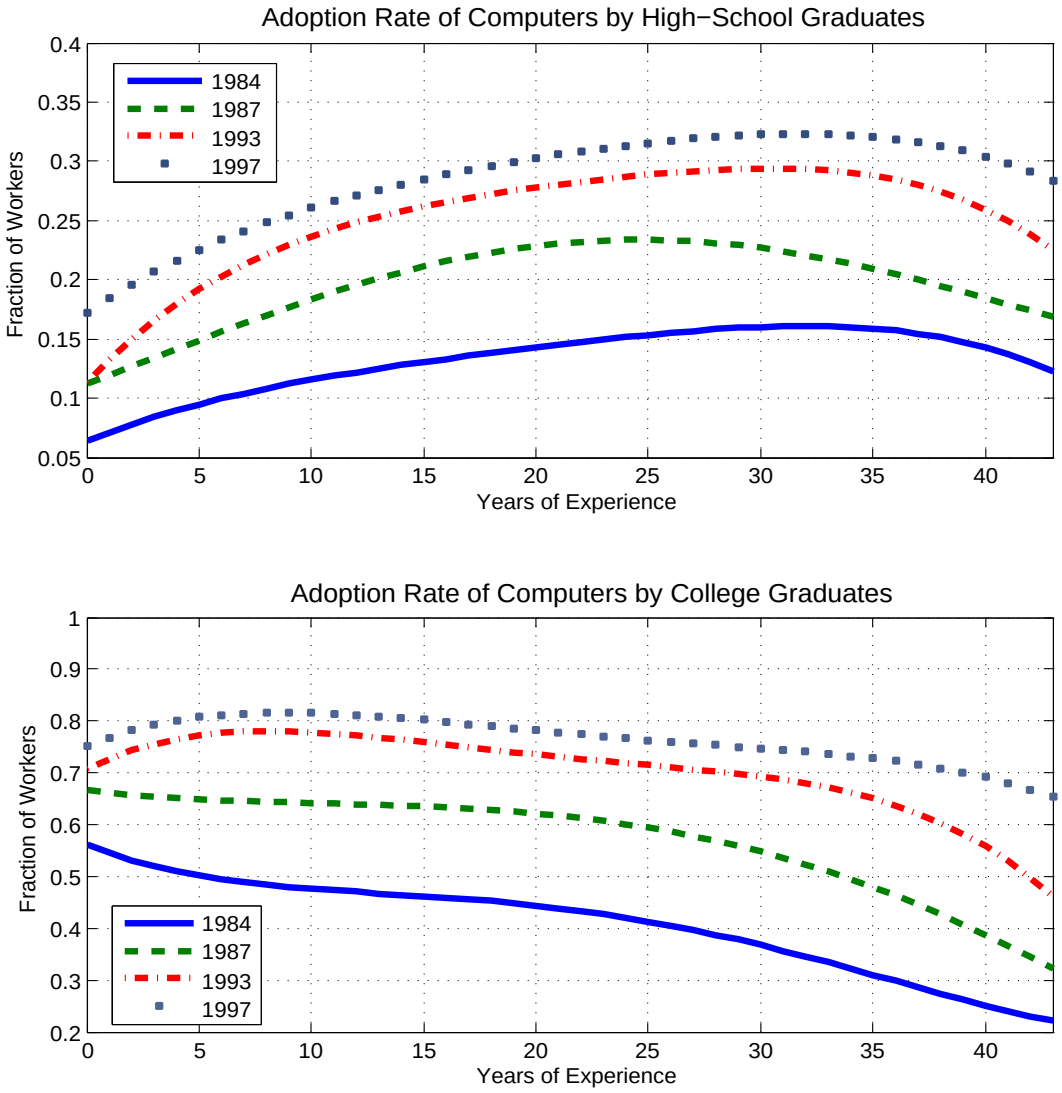


Figure 5: The top panel depicts the experience profile of the adoption rate of computers for U.S. high-school graduates for 1984, 1987, 1993, and 1997. The bottom panel plots similar experience profiles for college graduates. The figure is reproduced from Weinberg (2003b).

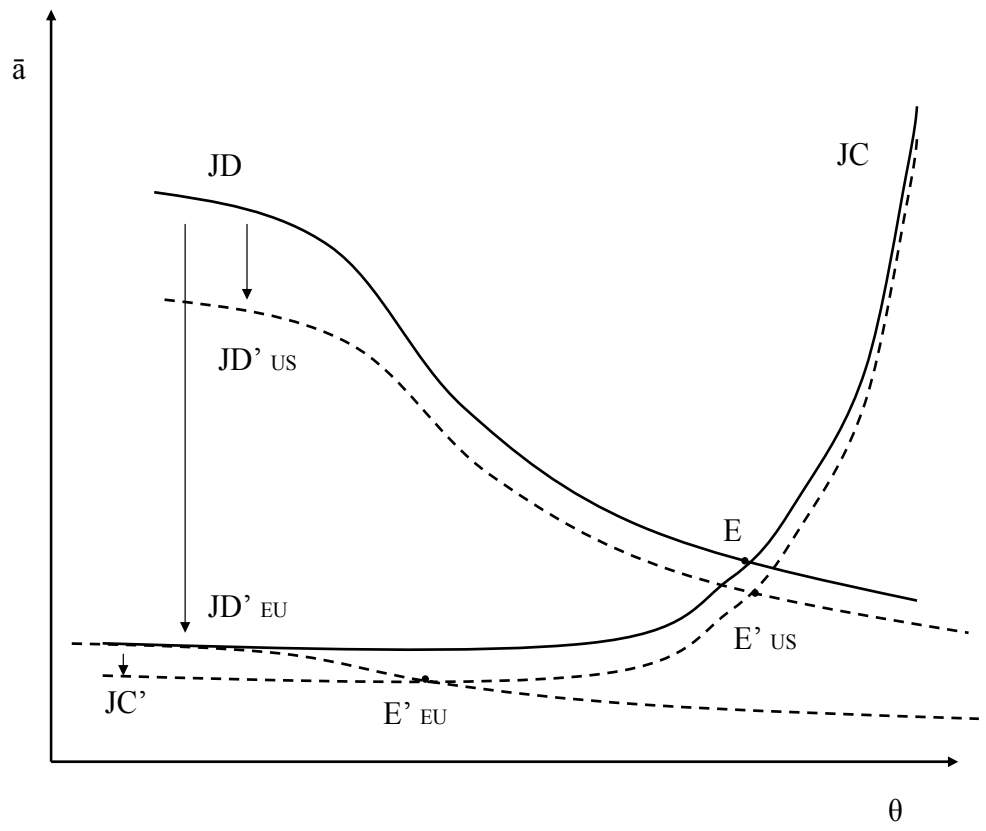


Figure 6: The figure illustrates graphically the equilibrium comparative statics of the model by Hornstein, Krusell and Violante (2003a). Following an acceleration in the rate of capital-embodied technical change, both the job-creation (JC) and the job-destruction (JD) curves shift. The amplitude of the shift is regulated by institutions, and hence it differs between the flexible economy (US) and the rigid economy (EU).

September 5, 2003
Preliminary

A Unified Theory of the Evolution of International Income Levels

Stephen L. Parente
and
Edward C. Prescott*

ABSTRACT: This paper develops a theory of the evolution of international income levels. In particular, it augments the Hansen-Prescott theory of economic development with the Parente-Prescott theory of relative efficiencies and shows that the unified theory accounts for the evolution of international income levels over the last millennium. The essence of this unified theory is that a country starts to experience sustained increases in its living standard when production efficiency reaches a critical point. Countries reach this critical level of efficiency at different dates not because they have access to different stocks of knowledge, but rather because they differ in the amount of society-imposed constraints on the technology choices of their citizenry.

* Parente, University of Illinois at Urbana-Champaign; Prescott, University of Minnesota and Federal Reserve Bank of Minneapolis. Prescott thanks the National Science Foundation and the University of Minnesota Foundation for research support. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

1. Introduction

Over the last decade, a fairly complete picture of the evolution of international income levels has emerged. Figure 1 plots the path of gross domestic product (GDP) per capita for four major regions of the world relative to the leader going back to 1700 using data from Maddison (1995). In 1700, the living standard of the richest country was less than three times the living standard of the poorest country.[1] This is the nature of the disparity prior to 1700 as well, as no single country experienced sustained increases in its living standard over the pre-1700 period. After 1700, huge differences in international incomes emerged, as some countries experienced sustained and large increases in their living standards well before others.

England was the first country to develop, that is, to realize sustained increases in per capita income. The exact date at which England began to develop is subject to debate. Some historians such as Bairoch (1993) place this date at around 1700. Western European countries and countries that were ethnic offshoots of England began to develop shortly thereafter. At first, the increases in income experienced by these early developers were irregular and modest in size. For example, Bairoch (1993) reports that it took England nearly 100 years to double its income from its 1750 level. However, after the start of the twentieth century, these increases have been larger and relatively regular with income doubling every 35 years in these countries—a phenomenon Kuznets (1966) labels *modern economic growth*.

Countries located in other regions of the world started the development process later in time. For these countries, the gap in income with the leader continued to widen

[1] Bairoch (1993) estimates this difference in 1700 to be smaller than a factor of two.

prior to the time they started modern economic growth. For Latin America, the beginning of the twentieth century is the approximate start of modern economic growth. For Asia, the middle of the twentieth century is the approximate start of modern economic growth. For Africa, modern economic growth has yet to start: although per capita income has increased in nearly every African country since 1960, the increases have been modest and irregular in the period that has followed. Because of these later starting dates, the disparity in international income levels increased to their current-day levels.

Some countries and regions have dramatically reduced their income gap with the leader subsequent to starting modern economic growth. For example, in the postwar period, Western Europe has managed to eliminate much of its income gap with the United States, the leader since 1890. Asia is another region that has been catching up with the leader in the postwar period. The catch-up in Asia, in fact, has been dramatic because of the growth miracle countries of Japan, South Korea, and Taiwan that doubled their income in a decade or less. Latin America, in contrast, is an example of a region that has not eliminated its gap with the leader since starting modern economic growth. Latin American per capita income has remained at roughly 25 percent of the leader for the last 100 years.

A theory of the evolution of international income levels must account for these facts. The theory must generate an initial period with living standards at the pre-1700 level followed by a long transition period to modern economic growth. The theory must generate different starting dates for the transition to modern economic growth across countries. Namely, it must identify some factor or set of factors that differs across countries and that delays the start of the transition by as much as two centuries. The

theory must also account for the sizable and persistent differences in living standards that characterize the experience of countries that have been experiencing modern economic growth for as long as 100 years. Finally, the theory must be consistent with *growth miracles*, namely, the large increases in relative income experienced by some initially poor countries in a relatively short period of time after 1950.

There are well-tested theories of some of these phenomena, but not a comprehensive theory that accounts for all of them. This paper unifies these well-tested theories and examines whether the unified theory can account for all of these phenomena.[2] A well-tested theory of the first phenomenon, the pattern of an initial period of stagnant living standards followed by a transition to modern economic growth, is provided by Hansen and Prescott (2002). The Hansen and Prescott theory is a combination of two long-standing and successful theories: the classical theory of the pre-1700 period and the neoclassical theory of the post-1900 period.

The classical economists, in particular, Malthus (1797) and Ricardo (1817), developed a theory that accounts well for the constant living standard that characterized the pre-1700 era. The main feature of this theory is an aggregate production function characterized by fixed factors, the most important of which is land. According to this theory, increases in knowledge lead to increases in output that are completely offset by increases in population. As a result, living standards do not increase. Economists have also had for a long time a good theory of modern economic growth that has characterized the United States and much of Western Europe since 1900. Solow (1970) developed his

[2] Ngai (2000) provides a unification of these theories along the lines of this paper.

growth model specifically to account for this post-1900 pattern of growth. The main feature of this theory is also an aggregate production function, but one with no fixed factor of production. According to this theory, improvements in technology that lead to more output being produced with the same resources are not offset by increases in population. As a result, living standards rise.

Hansen and Prescott (2002) unify the classical and modern growth theories by allowing people to use both the traditional production function and the modern production function. They show that when total factor productivity (TFP) associated with the modern production function reaches a critical level, the economy moves resources out of the traditional sector and into the modern sector. This is the date at which the transition begins. The transition is found to last a long period, roughly a century. The model thus gives rise to a pattern of economic development characterized by a long initial period of economic stagnation, followed by a long transition, followed by modern economic growth, as observed in Western Europe and its offshoots.

The Hansen and Prescott theory is not a theory of the evolution of international income levels because it does not address the issues of different starting dates of the transition to modern economic growth, sizeable income differences for countries experiencing modern economic growth, and growth miracles. Some factor that differs across countries must be added to the Hansen and Prescott theory to make it a theory of the evolution of international income levels.

Parente and Prescott (2000) develop a theory that accounts for the sizable differences in living standards for countries experiencing modern economic growth and that accounts for growth miracles. More specifically, they develop a theory of a country-

specific TFP and then introduce this factor into a model in which only the modern production function is available. Their theory of country-specific TFP, which they refer to as a *theory of relative efficiency*, is based on policy differences. More specifically, the theory shows how various policies that constrain choices of technology and work practices at the level of the production unit determine the aggregate efficiency at which a country uses its resources in production. The development of a theory of relative efficiencies is essential. Despite the fact that there is ample empirical evidence that countries differ in relative efficiencies, a theory of international income levels that takes countries' TFPs as exogenous is sterile, because it offers no policy guidance.

In this paper, we augment the Hansen and Prescott theory of economic development with the Parente and Prescott (2000) theory of relative efficiencies and show that the resulting unified theory is a theory of the evolution of international income levels. In this unified theory, a country begins its transition to modern economic growth when the efficiency with which it uses resources in the production of goods and services in the modern sector reaches a critical point. Countries reach this critical level of efficiency at different dates not because they have access to different stocks of knowledge, but rather because they differ in the amount of society-imposed constraints on the technology choices of their citizenry. We show that plausible differences in efficiencies delay the start of the transition to modern economic growth by more than two centuries, as observed in the data. Additionally, we show that the augmented model accounts well for the growth miracles that a number of countries experienced subsequent to 1950. Changes in a country's institutions that result in large increases in the efficiency

with which resources can be used in production give rise to growth miracles. Thus, the unified theory accounts for the way international income levels have evolved.

The paper is organized as follows. Section 2 starts with a review of the classical theory of the pre-1700 income level followed by a review of the neoclassical growth theory of modern economic growth. It then concludes with a review of how Hansen and Prescott (2002) combine these two theories into a single theory of economic development. Section 3 deals with the second component of the theory, namely, differences in efficiencies. It reviews the Parente and Prescott (2000) theory of relative efficiencies. Section 4 develops the unified theory of international income levels. In Section 4, a model based on the unified theory is developed and calibrated to the U.K. and U.S. development experiences over the last three centuries. The calibrated model is used to examine the effect of differences in efficiencies across countries on the start of the transition to modern economic growth and the effect of an increase in a country's efficiency on the subsequent path of its per capita GDP. Section 5 examines the development experiences of individual countries and groups of countries over the last three centuries within the context of the theory. Section 6 concludes the paper.

2. A Theory of Economic Development

In this section, we present the theory of economic development put forth by Hansen and Prescott (2002). We do this in three stages. First, we describe the classical component of that theory and derive its equilibrium properties. Next, we describe the modern growth component of that theory, and also derive its equilibrium properties. The last stage merges these two components and, in doing so, presents the Hansen and Prescott model of economic development.

Figure 2 describes the general pattern of economic development. More specifically, Figure 2 reports per capita income of the leader country dating back to 2000 B.C. Up until 1700, the living standard in the leader country, or any other country for that matter, displayed no secular increase. These living standards were significantly above the subsistence level. In 1688, for example, the poorest quarter of the population in England—the paupers and the cottagers—survived on a consumption level that was roughly one-fourth the national average.[3] A few societies, such as the Roman Empire in the first century, the Arab Caliphates in the tenth century, China in the eleventh century, and India in the seventeenth century realized some increase in their per capita income. However, these increases were not sustained. After 1700, per capita income in England started to increase. Over the next 150 years, these increases in the leader country were modest in size and irregular. However, since 1900, these increases have been larger and fairly regular, with per capita income doubling roughly every 35 years.

Technology was not stagnant over any part of this time period. Economic historians have documented a steady flow of technological innovations in this 2000 B.C. to A.D. 1700 period.[4] Yet these innovations prior to 1700 did not translate into increased living standards. Instead they translated into increased population: as total output increased, the population adjusted so as to maintain a constant level of per capita output. After 1700, these innovations did translate into increases in living standards.

[3] See Maddison (1991, p. 10) and Bairoch (1993, pp. 101–108).

[4] See Mokyr (1990) for a review of this literature.

2A. Classical Theory: The Pre-1700 Era

Classical economists, most notably Malthus and Ricardo, devised a theory that accounts well for the constant level of per capita income that characterized the pre-1700 era. The theory predicts a trade-off between living standards and population size. This trade-off exists because population growth is an increasing function of per capita consumption and because there is an important fixed factor of production, namely, land. A key implication of this theory is that there is a constant standard of living to which the economy adjusts. The theory predicts that increases in the stock of usable knowledge, which could translate into increases in living standards, instead translate into an increase in population.

Malthus' theory of population is a biological one rather than an economic one. According to his theory, fertility rises and mortality falls as consumption increases. Being classical, the model has no utility theory and so agents have no decision over the number of children they have. Recently a number of authors, including Tamura (1988), Becker, Murphy, and Tamura (1990), Doepke (2000), Galor and Weil (2000), and Lucas (2002), have generated Malthus-like population dynamics in a neoclassical model with household utility defined over consumption of goods and number of children. These models follow Becker (1960) by having a trade-off between quality and quantity of children.

We take an alternative approach, one that has society determining the size of its population through its institutions and policies. We likewise add household preferences to the classical theory of production. However, we define household preferences only over household consumption and not the quantity of children. Consequently, in this societal

theory of population growth, the quantity of children is treated as exogenous from the standpoint of the household.

The reason we take this societal approach to population growth is twofold. First, there is no tested theory of population dynamics, and once modern economic growth begins, demographics play a secondary role in development. Second, and more important, the approach reflects the view that groups of individuals, namely, societies, have had a much larger say in deciding how many children a family has than the family itself. Societies have instituted and continue to institute policies that give them their desired population size. Often the policies of society are not what individual families want. In modern China, for example, a law effectively limits many households to one child. By contrast, Iran in the 1980s wanted a higher population and so implemented subsidies to encourage people to have more children. After achieving its objective, the government stopped these subsidies in the 1990s and began to subsidize contraceptives. India today, wanting a lower population growth rate, has set up family planning programs in many regions. In all these cases, the effects of policy upon demographics are dramatic. Even in poor and rural Indian villages, which did not experience any increase in human capital or income, policy has led to a dramatic decline in population growth rates.

Why did society choose population size prior to 1700 so as to maintain the same constant living standard? The answer relates to the fact that land was an essential input to the production process in the pre-1700 era. In particular, as a valuable resource, land was subject to expropriation by outsiders. Prior to modern times, a small group of people with large amounts of quality-adjusted land and therefore a high income standard could not defend this land from outside expropriators. For this reason, there was a maximal

sustainable living standard. Society set up social institutions that controlled population so as to maintain the highest possible living standard consistent with the ability to defend itself from outside expropriators. Once an economy switches to the modern production technology, land is no longer an important input, so its defense is not an issue. At the stage when the modern production technology dominates, society sets up its social institutions that it sees as maximizing living standards subject to a constraint that a society perpetuates itself.

For the purpose at hand it is not essential that we model society's choices of institutions that affect fertility choices. Instead, it is sufficient to treat the growth rate of population in a simple mechanical way, namely, as a function of average consumption. In order to reflect society's choices, it must display two properties. First, the function must have a large slope, in the neighborhood of the pre-1700 consumption level. Second, for high levels of average consumption, the slope of the function must be near zero. The first property is only relevant for the theory of the pre-1700 era. The second property is only relevant for the theory of the post-1900 period. This is the approach that we take in this chapter.

With this in mind, we now proceed with a neoclassical formulation of the classical theory of constant living standards. There is a single good in the model that can be used for either consumption or investment purposes. The good is produced with a constant returns to scale technology that uses capital, labor, and land. An infinitely lived household owns the economy's land and capital and rents them to firms in the economy. Land is fixed and does not depreciate. The household is made up of many members, each of whom is endowed with one unit of time. The household uses its capital, labor, and land

income for consumption and investment purposes. The growth rate of population is a function of average consumption of household members. A household member's utility in the period is defined over the member's consumption in the period. The household's objective is to maximize the sum of each member's utility. The details of the economy are described as follows.

Technology

The classical theory of production is given by a Cobb-Douglas technology,

$$Y_{Mt} = A_{Mt}\, K_{Mt}^{\phi} N_{Mt}^{\mu} L_{Mt}^{1-\phi-\mu}. \tag{2.1}$$

In equation (2.1), Y_{Mt} is output, K_{Mt} is capital, N_{Mt} is labor, and L_{Mt} is land in period t. A_{Mt} is a total factor productivity (TFP) parameter, ϕ is the capital share parameter, and μ is the labor share parameter. The Cobb-Douglas assumption implies unit elasticity of substitution.[5] We allow for exogenous growth in TFP. More specifically, we assume that technology grows at the exogenous rate of γ_M; that is, $A_{Mt} = (1+\gamma_M)^t$. This assumption reflects the fact that technological change was evident from 2000 B.C.[6]

Output can be used for either consumption or investment purposes. The resource constraint for the economy is given by

$$C_t + X_t = Y_{Mt}, \tag{2.2}$$

where C_t denotes total consumption and X_t denotes total investment.

[5] The precise value of the elasticity of substitution between land and the other factors is not important provided that it is not greater than one. The evidence is that throughout most of history the substitution of these other factors for land was limited and, if anything, this elasticity of substitution was less than one. The unit elasticity assumption is made because it simplifies the analysis.

[6] We follow Hansen and Prescott's convention of using the letter *M* to index variables associated with the classical production function.

Preferences

Household preferences are added to the classical theory of production as follows. Period utility of each household member is defined over the member's consumption of the final good. We assume a log utility function, because it is in the class of utility functions that is consistent with a constant-growth equilibrium and because empirically it is consistent with a wide variety of micro and macro observations. Household utility in each period is the sum of each individual member's utility in the period. Strict concavity of individual household members' preferences implies that the household's utility is maximized by giving equal consumption to each member. For this reason, the discounted stream of utility of the household is just

$$\sum_{t=0}^{\infty} \beta^t N_t \log(c_t), \tag{2.3}$$

where β is the time discount factor, c_t is consumption of a household member, and N_t is household size.

As is evident from equation (2.3), we are using a dynastic construct. This is in contrast to Hansen and Prescott (2002), who use a two-period overlapping generations construct. We adopt an infinitely lived household framework rather than the two-period overlapping generations framework for two reasons. First and foremost, the empirical counterpart of a period is a year, while in the two-period overlapping generations construct, the empirical counterpart of a period is 35 years. Thirty-five years is simply too long a period for examining the model's ability to account for the large increases in output realized in a short period of time after 1950 by countries such as Japan and South Korea and for the long transition to modern economic growth.

Second, the size of the effect associated with differences in savings rates on an economy's steady-state per capita output level depends importantly on the construct that is used. The level effects are in fact larger with the dynastic construct. This is important for judging whether differences in savings rates can account for the large differences in transition dates as well as the large differences in incomes that continue to exist between economies that have started modern economic growth. Thus, if plausible differences in savings rates cannot give rise to 200-year delays in development in the dynastic construct, then it follows that some factor other than savings rates accounts for the pattern of development.[7] This is the conclusion of the quantitative exercises undertaken by Parente and Prescott (2000). The choice of construct is not important, however, in assessing the plausibility of other factors such as efficiency, as reflected in TFP, differences: the size of the level effects is the same regardless of whether the dynastic or overlapping generations construct is employed.

Endowments

Each member of the household is endowed with one unit of time, which the member can supply to firms in the economy to earn wage income. The household is also endowed with the economy's stock of land and capital, which the household rents to firms in the economy. Land in the economy is fixed in supply: it cannot be produced, and it does not depreciate. Without loss of generality, the total quantity of land in the economy is normalized to one. Since land has no alternative use aside from production, the input to production in each period is one. Capital is assumed to depreciate and evolves according to the following law of motion:

[7] See Hendricks (forthcoming) for a more detailed explanation of this phenomenon.

(2.4) $$K_{t+1} = (1-\delta)K_t + X_t,$$

where δ is the depreciation rate.

Population Dynamics

As mentioned earlier, because we take a societal approach to population size, we model population growth as a function of the average consumption level of household members. More specifically, we assume that the number of agents born into a household in period $t + 1$ depends on the average consumption level of household members from period t. Let N_t denote the number of household members in period t, and let c_t denote their average consumption level. Then,

(2.5) $$N_{t+1} = g(c_t)N_t.$$

The function g is the growth factor of population from one period to the next. The classical prediction of a stable living standard at the pre-1700 level, c_M, requires that the function g have a sufficiently large and positive slope at c_M and that $g(c_M) = (1+\gamma_M)^{1/(1-\phi-\mu)}$. This c_M is the maximal living standard consistent with a society being able to defend its land.

Equilibrium Properties

For such a population growth function, there is a steady-state equilibrium with a constant living standard c_M and a population growth rate equal to $(1+\gamma_M)^{1/(1-\phi-\mu)} - 1$. This constant living standard satisfies $g(c_M) = (1+\gamma_M)^{1/(1-\phi-\mu)}$. Were the living standard to rise above c_M, say, because of plague or drought, population increase would exceed technical advances and the living standard would then fall until it returned to c_M. If for some reason c were below c_M, the population growth factor would be less than the one needed to

maintain the living standard, and the living standard would increase until it was again c_M. Along the steady-state equilibrium path, aggregate output, capital, consumption, and the rental rate of land all grow at the rate of the population. Per capita variables as well as the rental price of labor and capital are all constant. Increases in technology in this model simply translate into a higher population rather than higher living standards. This is precisely the pattern of development observed prior to 1700.

2B. Modern Growth Theory: The Post-1900 Era

The classical theory accounts well for the pattern of economic development up to 1700. However, it does not account for the increase in living standards that occurred after 1900. Since about 1900, the growth rate has been roughly constant, with a doubling of per capita output every 35 years. Modern growth theory, in contrast, does. We now turn to that theory.

Besides the roughly constant rate of growth achieved by developed countries over the last century and a half, a number of other features of post-1900 growth in the United Kingdom and some other countries are noted by Kaldor (1957). These additional modern economic growth facts are roughly that the consumption and investment shares of output are constant, the share of income paid to capital is constant, the capital-to-output ratio is constant, and the real return to capital is constant.

Modern growth theory accounts well for these modern growth facts. Quantitatively, the steady-state equilibrium of the economy mimics the long-run observations of the United Kingdom and the United States. This is no surprise: Solow (1970) developed the theory with these facts in mind. A key feature of that theory is a Cobb-Douglas production function that includes no fixed factor of production and that is

subject to constant exogenous technological change. More specifically, the production technology for the composite good that can be used for either consumption or investment purposes is given by

$$Y_{St} = A_{St}\ K_{St}^{\theta}\ N_{St}^{1-\theta}\ . \tag{2.6}$$

In equation (2.6), Y_{St} is output, K_{St} is capital, and N_{St} is labor in period t. The parameter θ is capital's share, and the parameter A_{St} is TFP. TFP grows exogenously at the constant, geometric rate γ_S. As can be seen, the critical difference between the classical and modern growth production functions is that the modern growth function does not include the fixed factor input land.[8]

Because the final objective of this section is to merge the classical theory and the modern growth theory into a single model, we maintain the same assumptions regarding preferences, endowments, and population dynamics as in the preceding subsection. The household in the model rents capital to firms and supplies labor. It uses its capital and labor income to buy consumption for household members and to augment the household's stock of capital.

In contrast to the classical theory, population growth in the modern theory does not have any consequences for the growth rate of per capita variables in the long run. The choice of the population growth function is therefore unimportant in this respect. The standard procedure is to assume a population growth function $g(c)$ that is constant over the range of sufficiently high living standards associated with the modern growth era. Population thus grows at a constant exponential rate.

[8] Again, we follow Hansen and Prescott's convention of using S to index variables associated with the modern growth production function.

Clearly, population cannot grow at an exponential rate forever. At some population level, natural resources would become a constraining factor. If population were ever to reach this level, it would be unreasonable to abstract from land as a factor of production. But societies control their population so that it never reaches this level. Indeed, reproduction rates have fallen dramatically in the last 50 years, so much in the rich countries, in fact, that these countries must increase their fertility rates to maintain their population size in the long run. This suggests a population growth function that asymptotically approaches one. This is an additional property we impose on the population growth function in the analysis that follows.

In the case where the population growth function is a constant, per capita output, consumption, and capital all increase at the rate $(1+\gamma_S)^{1/(1-\theta)}$ along the equilibrium constant growth path. The rental price of labor also grows at this rate. The rental price of capital, in contrast, is constant. Capital's share of income is also constant and equal to θ, as is consumption's share and investment's share of output. As can be seen, the growth rate of the economy's living standard is independent of the economy's population growth rate: the only thing that matters is the exogenous growth rate of technological change. The population growth rate does have a level effect, but it is small. Thus, unlike in the model of the pre-1700 era, the population growth function in the model of the post-1900 era has only a minor role.

2C. The Combined Theory

The classical theory accounts well for the constant living standard that characterizes the pre-1700 era, and the modern growth theory accounts well for the doubling of living standards every 35 years that characterizes the post-1900 experience of

most of the currently rich, large, industrialized countries. In the period in between, living standards increased in these countries, but at a slower and far more irregular rate compared to the post-1900 period.

We seek a theory of this development process, namely, a theory that generates a long period of stagnant living standards up to 1700, followed by a long transition, followed by modern economic growth. Given the success of the classical theory and the modern growth theory in accounting for the pre-1700 and post-1900 eras, the logical step, and the one taken by Hansen and Prescott (2002), is to merge the two theories by permitting both technologies to be used in both periods. We now present the combined theory of Hansen and Prescott, and we use that theory to organize and interpret the development path of the leading industrialized country over the 1700–2000 period.

In the combined theory of Hansen and Prescott (2002), output in any period can be produced using the traditional and/or the modern growth production functions. Both technologies, therefore, are available for firms to use in all periods.[9] Capital and labor are not specific to either production function. In light of these assumptions, the aggregate resource constraint for the combined model economy is

$$N_t c_t + X_t \leq Y_{Mt} + Y_{St} = Y_t, \tag{2.7}$$

the capital rental market clearing constraint is

$$K_t = K_{Mt} + K_{St}, \tag{2.8}$$

and the labor market clearing condition is

[9] The maximum output that can be produced if both technologies are available is characterized by a standard aggregate production function $Y_t = A_t F(K_t, L_t, N_t)$. By *standard* we mean that it is weakly increasing and concave, homogenous of degree one, and continuous. Even though both the Malthus and the Solow production functions are Cobb-Douglas technologies, the function *F* is not Cobb-Douglas.

(2.9) $$N_t = N_{Mt} + N_{St}.$$

Household preferences continue to be given by equation (2.3). Additionally, the population growth function continues to be given by equation (2.5), and it displays the properties that the function has a large slope in the neighborhood of the pre-1700 consumption level and a slope near zero for large levels of consumption.

In their combined theory, Hansen and Prescott assume that the rate of TFP for the classical production function and the rate of TFP for the modern economic growth production function are each constant over time. We deviate from Hansen and Prescott on this dimension. Although we maintain their assumption that the rate of TFP growth associated with the traditional technology is constant, we assume that the rate of TFP growth associated with the modern growth technology increases over time, converging asymptotically to the modern growth rate. We make this alternative assumption in light of the historical evidence on technological change and the empirical counterparts of the two production functions.

The empirical counterpart of the classical production function is a traditional technology for producing goods and services that is most commonly associated with the family farm. A key feature of this production technology is that it is based on the use of land in the production of hand tools and organic energy sources. For this technology, the historical record shows gradual improvements in these methods over the last 2,000 years at a roughly constant rate change.[10] The empirical counterpart of the modern growth production function is a modern technology that is most commonly associated with the

[10] The exception to this constant rate of growth might be the Green Revolution in the middle of the twentieth century, where the introduction of new seed varieties resulted in large increases in farm yields associated with traditional farming methods.

factory.[11] A key feature of this technology is that it uses machines driven by inanimate sources of energy. For this technology, the historical record suggests modest growth in the eighteenth century, followed by much higher growth in the nineteenth and twentieth centuries. Consequently, a more plausible assumption is that the growth of TFP associated with the modern production function increased slowly after 1700 and converged to the rate associated with the modern growth era shortly after 1900.

We emphasize that traditional production occurs in household production units with most of the resources being allocated to producing household consumption and only a limited amount to trade. There was little scope for people working in these sectors to develop more efficient production methods. Rapid increases in productivity occurred only when goods developed in the industrial sector were introduced in farming. The reaper and the tractor dramatically increased productivity on farms. Insecticides and fertilizers also contributed to productivity, as did the development of hybrid corn and new seeds. This is all well-documented by Johnson (2000).

An economy that starts out using only the traditional production function will eventually use the modern one. To see this, suppose that it were never profitable for firms to use the modern production function. Then the economy's equilibrium path would converge to the steady state of the pre-1700-only model. The steady state of that model is characterized by constant rental prices for capital and labor, r_M and w_M. Capital and labor are not specific to any one technology. Thus, a firm that first considers using the modern production function can hire any amount of capital and labor at the factor rental prices r_{Mt}

[11] The distinction between technologies is, thus, not along the lines of agriculture and manufactures. In this classification, modern agriculture with its use of synthetic fertilizers and tractors is associated with the modern growth production function.

and w_{Mt}. Profit maximization implies that a firm will not choose to operate the modern growth technology if

$$A_{St} < \left(\frac{r_{Mt}}{\theta}\right)^{\theta}\left(\frac{w_{Mt}}{1-\theta}\right)^{1-\theta}. \tag{2.10}$$

This inequality must be violated at some date. Asymptotically, the rental prices would approach constant values if only the classical production function were operated, and so the right-hand side of (2.10) is bounded. The left-hand side is unbounded because TFP in the modern function grows forever at a rate bounded uniformly away from zero. The inequality given by (2.10), therefore, must be eventually violated. At the date when TFP in the modern production function surpasses the critical level given by the right-hand side of (2.10), the economy will start using the modern growth production function. This marks the beginning of the Industrial Revolution. This result is independent of the size differences in the growth rates of TFP associated with the traditional and modern production functions.

Over the transition, more and more capital and labor will be moved to the modern production sector. The rental price of labor will show a secular rise. The traditional production function will, however, continue to be operated, though its share of output will decline to zero over time, because of the assumptions that land is used only in traditional production and that its supply is inelastic.

We now use the combined theory to organize and interpret the development path of the industrial leader over the 1700–2000 period. The empirical counterpart of a period is a year. The initial period of the model is identified with the year 1675. We attribute the stagnation of the leader prior to 1700 to a low level of TFP associated with the modern production function to warrant use of the modern production function. We attribute the

start of economic growth of the leader in 1700 to growth in TFP associated with the modern production function so that its level exceeds the critical value given by equation (2.10). Lastly, we attribute the rising rate of growth of per capita output of the leader from 1700 to 1900 to greater use of the modern production function and the rising rate of growth of TFP.

We proceed to parameterize the model. The model is calibrated so that the economy starts to use the modern production function around the year 1700. Following Hansen and Prescott, the model is calibrated so that the steady state of the classical-only model (subsection 2A) matches pre-1700 observations and the steady state of the modern growth-only model (subsection 2B) matches the post-1900 growth experience of the United States.

In the calibration, we deviate from Hansen and Prescott along two dimensions. First, we calibrate the population growth function so that it matches Maddison's (1995) estimates for U.K. population growth rates over subperiods of the 1675–1990 period. Given our theory of population growth, it is more appropriate to use the time series data from a particular country to restrict the population growth function for that country rather than cross-section data as Hansen and Prescott do. Second, we calibrate the annual growth rate of TFP for the modern production function so that it remains at the traditional rate up until 1700, increases linearly to one-half of its modern growth rate in 1825, and then increases linearly to its modern growth rate in 1925.

Following Hansen and Prescott, we pick the initial capital stock and the initial population so that if only the traditional production function were available, the equilibrium would correspond to the steady state of the pre-1700 model and there would

be no incentive to operate the modern production function if it were available. This ensures that in period 0 only the traditional production function is operated and that there is a period of constant living standards.

Table 1 lists the values for each of the model parameters and provides comments where appropriate. The population growth rate function implied by the U.K. population growth data used in the computation is depicted in Figure 3.

For the parameterized model economy, it takes 150 years before 95 percent of the economy's output is produced in the modern sector. Figures 4–6 depict the model economy's development path along a number of other dimensions. Figure 4 compares period *t* per capita output relative to 1700 per capita output for the model economy and the industrial leader as reported by Maddison (1995, Tables 1.1 and C.12). According to the model, an economy that begins the transition in 1700 will be approximately 28 times richer in 1990 as it was in 1700. Figure 5 depicts the growth rate of per capita output for the model economy over the 1700–2000 period. The growth rate of per capita output is slow at the onset of the transition, less than 1 percent per year on average. One hundred years later, the growth rate is near the modern growth rate of 2 percent per year. This pattern is primarily a consequence of the assumption that TFP growth for the modern technology increases slowly over the 1700–1990 period. Figure 6 depicts the path of the rental prices of capital and labor over the 1700–2000 period. As can be seen, the real wage rate increases steadily once the transition begins. The real interest rate, in contrast, shows very little secular change over three centuries. These latter predictions conform well to the pattern of development associated with England, the United States, and other early developers.

Table 1. Restricted Parameter Values

Parameter	Value	Comment
γ_M–growth rate of TFP for traditional production	.0009	Consistent with pre-1700 world population average annual growth rate of .003
ϕ–capital share in traditional production	.10	
μ–labor share in traditional production	.60	Chosen so that labor's share does not vary with the level of development as reported by Gollin (2002)
A_{M0}–initial TFP for traditional production	1.0	Normalization
δ–depreciation rate	.06	Consistent with U.S. capital stock and investment rate since 1900
γ_S–asymptotic TFP growth rate for modern production	.012	2 percent rate of growth of per capita GDP in modern growth era
θ–capital's share in modern production	.40	U.S. physical capital's share of output
A_{S0}–initial TFP for modern production	.53	1700 starting date given initial period for model is 1675
β–subjective time discount factor	.97	Consistent with real rate of interest between 4 and 5 percent in modern growth era

The predictions of the model are not sensitive to the value of the capital share parameter in the modern growth production function. This is an important result, because the magnitude of the capital share with a broad definition of capital that includes intangible as well as tangible capital could well be greater than the 0.40 share value used in the above exercise. The paths of per capita GDP, its growth rate, and rental prices are nearly identical to those shown in Figures 4–6 for alternative values of the capital share in the modern production function. The transition still takes a long time. For a capital share as high as 0.70, 140 years elapse before 95 percent of the economy's capital is produced using the modern production function.

3. A Theory of Relative Efficiencies

The Hansen and Prescott theory of economic development reviewed in Section 2 is not a theory of the evolution of international income levels. It does not address the issue of why modern economic growth started at different dates in different countries. India, for example, began modern economic growth nearly 200 years later than did the United Kingdom. As a result, India's income level relative to the leader fell from 50 percent in 1770 to only 5 percent in 1970. Neither does the theory address the issue of why some countries that have been experiencing modern economic growth for a century have failed to narrow the income gap with the industrial leader. Latin America, for example has remained at roughly 25 percent the U.S. income level since the second half of the nineteenth century when modern economic growth began there. The theory does not address the issue of why some countries in the 1950–2000 period have been able to substantially narrow the income gap with the industrial leader. These countries include Italy, Japan, Korea, Spain, and Taiwan, and all of which experienced a growth miracle. Some factor that differs across countries must be added to the Hansen and Prescott theory to make it a theory of the evolution of international income levels.

One might be led to introduce differences in TFP associated with the modern production to the model, because the Hansen and Prescott theory of development predicts that per capita income in a country starts to increase once TFP in the modern sector reaches a critical level. Moreover, there is ample evidence that countries (at least those experiencing modern economic growth) differ along this dimension.[12] Although it would be easy to introduce such differences into the Hansen and Prescott theory, it would not be

[12] See, for example, Klenow and Rodriguez-Clare (1997), Hall and Jones (1999), and Hendricks (2002).

useful, as long as country-specific TFP differences are treated exogenously. Absent a theory of the country-specific TFP component, the theory of the evolution of international income levels is sterile because it offers no policy guidance. What is needed is a policy-based theory of why TFP differs across countries at a point in time.

Parente and Prescott (2000) develop a theory of TFP that attributes differences in TFP to country-specific policies that both directly and indirectly constrain the choice of production units. Their theory of TFP is more appropriately called a *theory of relative efficiencies*. This is because Parente and Prescott (2000) decompose a country's TFP into the product of two components. The first component is a pure knowledge or technology component, denoted by *A*. The second is an efficiency component, denoted by *E*. In the context of the Hansen and Prescott model, the modern growth production function is

$$Y_{St} = E_S A_{St} \; K_{St}^{\theta} \; N_{St}^{1-\theta} . \tag{3.1}$$

The technology component of TFP, A_{St}, is common across countries. It is the same across countries because the stock of productive knowledge that is available for a country to use does not differ across countries.[13] The efficiency component differs across countries as the result of differences in economic policies and institutions. Here we consider the case in which a country's economic policies and institutions do not change, so E_s is not subscripted by *t*. The efficiency component is a number in the (0,1] interval. An efficiency level less than one implies that a country operates inside the production possibilities frontier, whereas an efficiency level equal to one implies that a country

[13] Much of the stock of productive knowledge is public information, and even proprietary information can be accessed by a country through licensing agreements or foreign direct investment.

operates on the production possibility frontier. Differences in efficiency, therefore, imply differences in TFP.

Relative efficiencies at a point in time, and not absolute efficiencies, can be determined using the production function and the data on quantities of the inputs and the output. Thus, it is not possible to determine if any country has an efficiency level equal to one, although we tend to doubt that this is the case. Changes in relative efficiencies of a given country can also be determined conditional on an assumption on the behavior of the technology component of TFP such as that it grows at some constant rate.

We now present the Parente and Prescott (2000) theory of relative efficiencies. To keep the analysis manageable, we present the theory of relative efficiencies in the context of an economy in which only the modern production function is available. The theory constitutes a theory of the aggregate production function when there are constraints at the production unit level. In light of this, we first review the theory underlying the aggregate production function. We then show how policy constraints give rise to an aggregate production function with a different efficiency level. We follow this by providing estimates of cross-country relative efficiencies associated with the modern production function using the mapping from policy to aggregate efficiency derived in this section with estimates of the costs imposed by a country-specific policy. Finally, we conclude this section with a discussion of why constraints on the behavior of the production units exist.

The Aggregate Production Function

Before developing the mapping from policy to aggregate efficiency, we briefly review the theory of the aggregate production function associated with modern growth.

The theory underlying the aggregate production function is as follows. In each period, there is a set of plant technologies *B*. A plant technology $b \in B$ is a triplet that gives the plant's output y_b and its capital and labor inputs, k_b and n_b. A plan $\{\lambda_b\}$ specifies the measure of every type of plant operated. The aggregate production function, that is, the maximum *Y* that can be produced given aggregate inputs *K* and *N*, is

$$Y = F(K, N) = \max_{\lambda \geq 0} \sum_b \lambda_b y_b \tag{3.2}$$

subject to the two resource constraints

$$\sum_b \lambda_b k_b \leq K \tag{3.3}$$

$$\sum_b \lambda_b n_b \leq N\,. \tag{3.4}$$

Assuming that this program has a solution, which it will under reasonable economic conditions, the aggregate production function will be weakly increasing, weakly concave, homogeneous of degree one, and continuous.

Empirically, the Cobb-Douglas aggregate production function is the one consistent with the post-1850 modern economic growth era. The question then is, What set of technologies *B* gives rise to the Cobb-Douglas aggregate production function? One such set is the set of plant technologies defined by

$$y \leq d(n) k^{\theta}\,. \tag{3.5}$$

The function $d(n)$ is an increasing and continuous function of the labor input. Assuming that $n^* = \arg\max d(n)\, n^{\theta-1}$ exists, the aggregate production function is

$$Y = A\, K^{\theta} N^{1-\theta}, \tag{3.6}$$

where $A = \max d(n)\, n^{\theta-1}$. With the assumption that the function d increases over time, the expression A will increase over time.

Consequences of Constraints for Aggregate Efficiency

Next, consider the plant production technology with constraints imposed on it. We consider two types of policy. The first type constrains how a particular plant technology can be operated. The second type constrains the choice of the production units that can be operated. For sure, a number of other types of policy have a similar effect, but they are not considered by Parente and Prescott (2000).[14]

The first type constrains how a given technology is operated. A policy that gives rise to this type of constraint is a work rule, which dictates the minimum number of workers or machines needed to operate a plant technology. In particular, suppose constraints are such that the input to a $b = (k,n,y)$ type plant must be $\phi_K k_b$ and $\phi_N n_b$ for all plant types where ϕ_K and ϕ_N exceed one. This implies that a particular technology, if operated, must be operated with excessive capital and labor. With these constraints, the aggregate production function is

$$Y = \phi_N^{\theta-1}\phi_K^{-\theta} A\ K^{\theta} N^{1-\theta} = E_S\ A\ K^{\theta} N^{1-\theta}, \tag{3.7}$$

where $E_S \equiv \phi_K^{-\theta}\phi_N^{\theta-1}$. This is the aggregate production function used in Section 2. If the nature of the constraints were to double the capital and labor requirements, then the efficiency measure would be one-half. If the nature of constraints is to quadruple both the capital and labor requirements, then the efficiency measure would be one-fourth.

[14] For example, Schmitz (2001) suggests a mapping of government subsidies to state-owned enterprises and aggregate efficiency.

The second type of policy constrains the choice of the production units that can be operated. This type of constraint can map into the efficiency parameter of an aggregate production with a composite capital stock made up of both physical and intangible components. Any policy that serves to increase the amount of resources the production unit must spend in order to adopt a better technology is a constraint of this nature. Such policies and practices take the form of regulation, bribes, and even severance packages to factor suppliers whose services are eliminated or reduced when a switch to a more productive technology is made. In some instances, the policy is in the form of a law that specifically prohibits the use of a particular technology. The empirical evidence suggests that this second type of constraint is more prevalent than the first.[15]

Following Parente and Prescott (2000), let the output of a quality *b* plant be given by the following equation:

$$y_t = b\, k_{Pt}^{\theta_P} [\min(n_t, \bar{n})]^{\theta_n} \qquad \bar{n} > 0, \quad \theta_P < 1. \tag{3.8}$$

With this technology, a minimum number of workers, $\bar{n}$, is required to operate a plant. The variable k_P denotes the physical capital input. The subscript *P* is introduced in order to differentiate physical capital from intangible capital. There are no increasing returns to scale in the economy, because if the inputs of the economy are doubled, the number of plants doubles.[16]

A plant's quality is a choice variable. To improve its quality, resources are needed. This resource cost is the product of two components. The first component is technological in nature and reflects the cost in the absence of constraints. The second

[15] See Parente and Prescott (2000) for a survey of this evidence.

[16] See Hornstein and Prescott (1993) for a detailed coverage of this technology.

component, denoted by ϕ_I, reflects the constraint itself. The function that gives the required resources a plant must expend to advance its quality from b to b' is

$$x_{bb'} = \phi_I \int_b^{b'} \left(\frac{s}{W_t} \right)^{\alpha} ds. \tag{3.9}$$

W_t is the stock of pure knowledge in the world in period t. Its growth rate is exogenous and equal to γ_W. Thus,

$$W_{t+1} = W_0 (1 + \gamma_W)^t. \tag{3.10}$$

Integrating (3.9) yields

$$x_{It} = \phi_I \frac{b_{t+1}^{\alpha+1} - b_t^{\alpha+1}}{W_0^{\alpha} (1+\gamma)^{\alpha t} (1+\alpha)}.$$

Let

$$k_{It} = \frac{\phi_I b_t^{\alpha+1}}{W_0^{\alpha} (1+\gamma)^{\alpha(t-1)} (1+\alpha)}.$$

The plant technology is specified by

$$y_t = \mu \phi^{-\theta_I} (1+\gamma)^t k_{It}^{\theta_I} k_{Pt}^{\theta_P} [\min(\bar{n}, n_t)]^{1-\theta_n}, \tag{3.11}$$

with

$$k_{It+1} = (1 - \delta_I) k_{It} + x_{It}, \tag{3.12}$$

where δ_I and μ are functions of α, γ, ϕ_I, and W_0 and $\theta_I = 1 - \theta_P - \theta_n$. The variable k_{It} has the interpretation of the plant's intangible capital stock, as it is the value of the plant's past investments in quality improvements. The sum of θ_I and θ_P is strictly less than one, so there is an optimal plant size.

Aggregating over plants implies the following equilibrium aggregate production relation:

(3.13) $$Y_t = E_S A_0 (1+\gamma_S)^t K_{It}^{\theta_I} K_{Pt}^{\theta_P} N_t^{1-\theta_P-\theta_I},$$

with $E_S \equiv \phi^{-\theta_Z}$. The laws of motion for the aggregate capital stocks are

(3.14) $$K_{I,t+1} = (1-\delta_I)K_{It} + X_{It}$$

(3.15) $$K_{P,t+1} = (1-\delta_P)K_{Pt} + X_{Pt}.$$

Now if the intangible capital stock has the same depreciation rate as physical capital, then the aggregate production relation with two capital stocks given by (3.13) maps into an aggregate production function with one capital stock and a large value for the capital share.[17] More specifically, the capital share in the one-capital-stock model, θ, is the sum of θ_I and θ_P. The relations between the two capital stocks, K_I and K_P, and the composite capital stock, K, are

(3.16) $$K_{It} = \frac{\theta_I}{\theta} K_t$$

(3.17) $$K_{Pt} = \frac{\theta_P}{\theta} K_t.$$

In the experiments that follow, this effectively is the underlying aggregate production function for the modern sector in the combined development theory of Hansen and Prescott when we consider capital share values greater than 0.40.[18]

[17] This requires an assumption that there is an additional resource cost associated with maintaining the plant's current quality. Such a cost could reflect, among other things, training for young workers who replace old workers retiring in the previous period.

[18] We say *effectively* because there are two technical issues in the combined theory when capital is broadly defined. First, if intangible capital is not an input into the traditional production function, then the economy will need to make some investments specifically in intangible capital prior to switching to the modern production function. Second, after the transition, as new plants open, they will have a lower technology level compared to older plants.

Estimates of Aggregate Relative Efficiency

The mappings developed in the preceding subsection allow us to impute the aggregate relative efficiency associated with the modern production function for various constraints. In general, the size of the effect of the constraint on a country's aggregate efficiency depends on the factor input affected by the constraint and on that input's share in the production function. In the special case where the constraints affect all inputs equally, that is, $\phi = \phi_n = \phi_I = \phi_P$, the individual factor shares are unimportant and the efficiency level of a country is just $E_s = \phi$. Hence, the implied difference in relative efficiencies is equal to the implied cost differences of policy. Thus, if the cost difference in policies between two countries is a factor of five, the implied factor difference in aggregate relative efficiency is also five.

Are factor differences in relative efficiency greater than five reasonable? Obviously, it is not possible to answer this question definitively without a comprehensive international study of the total costs of the constraints imposed by society. Some estimates of the cost differences associated with some country-specific policies do exist. Studies that estimate the costs of certain policies of individual countries that affect the technology and work practice choices of the production units located there do find that these costs vary systematically with income levels, with large differences existing between rich and poor countries. These studies suggest that factor differences in relative efficiencies could be easily as great as five.

For example, Djankov et al. (2002) calculate the costs associated with the legal requirements in 75 countries that an entrepreneur must meet in order to start a business. They find that the number of procedures required to start up a firm varies from a low of 2

in Canada to a high of 20 in Bolivia and that the minimum official time required to complete these procedures ranges from a low of 2 days in Canada to a high of 174 days in Mozambique. These costs do not reflect any unofficial costs involved with starting a firm, such as bribes, or bureaucratic delays. Because these official cost measures are positively correlated with indexes that incorporate measures of bribes, the true difference in start-up costs between low-cost and high-cost countries is surely even larger than those reported in the study.

Reasons for Constraints

The evidence strongly suggests that production units in poor countries are severely constrained in their choices, and the costs associated with these constraints are large. This prompts the question, Why does a society impose these constraints? A large number of studies, some of which are surveyed in Parente and Prescott (2000), suggest that constraints typically are imposed on firms in order to protect the interests of factor suppliers to the current production process. These groups stand to lose in the form of reduced earnings if new technology is introduced. These losses occur because either the input they supply is specialized with respect to the current production process or the monopoly power granted to them over the supply of a particular input is eroded.[19]

4. A Unified Theory of the Evolution of International Incomes

In this section we unify the Parente and Prescott (2000) theory of relative efficiencies and the Hansen and Prescott (2002) theory of development. The unified

[19] Parente and Prescott (1999) show in a model with no capital how a monopoly right granted to factor suppliers can significantly lower a country's efficiency. Herrendorf and Teixeira (2003) extend this model to include physical capital and show that these monopoly rights have even larger effects on a country's efficiency.

theory is then used to organize and interpret the evolution of international income levels. We unify the Parente and Prescott (2000) theory and the Hansen and Prescott theory as follows. We assume that technological increases in both sectors result from growth in world knowledge. Consequently, the technology component of TFP in each production function is the same across countries at any point in time. The paths for the technology components of TFP are determined as in subsection 2C by requiring that the leader country with an efficiency parameter in the modern sector set to one start its transition to modern economic growth in 1700. We then introduce differences in this efficiency parameter across countries. Given a country's relative efficiency parameter and the common path of the technology components of the TFPs, we compute the equilibrium path of the economy.

As mentioned in Section 3, we doubt than any country has or had an efficiency parameter equal to one. The assumption that efficiency in the leader is one in the unified theory is not important to any of the results as it is just a normalization. Again, only relative efficiencies matter and can be determined. This is the case for countries at a given time and across time in a given country.

We do not introduce cross-country differences in the efficiency parameter associated with traditional production. As mentioned in the introduction, incomes did differ slightly prior to 1700, with the richest countries being no more than two or three times richer than the poorest. One possible explanation for these pre-1700 differences in income levels is that countries differed in policies that increased the inputs required for producing goods with the traditional production function. Because this technology corresponds to traditional farming and even manufactures produced within a home

setting, we think the effect of policy differences for relative efficiencies associated with traditional production is small. For this reason, we favor the alternative explanation that some countries were better able to defend themselves from outside expropriations because of geography and thus were able to maintain a higher constant living standard during the pre-1700 era. Countries that enjoyed such an advantage were England and Japan.

We interpret delays in the start of the transition to modern economic growth to late starters having a lower relative efficiency in the modern sector, at least up until the date their transitions began. We attribute the persistent percentage between a country that started modern economic growth later than did the industrial leader to the continuation of its low relative efficiency. Finally, we attribute catch-up, including growth miracles, to large increases in relative efficiency in countries.

We begin by computing the relative efficiency of a late starter required to delay the start of its transition by a given length of time. The size of the required efficiency difference between the leader and the laggard that gives rise to any given delay is a function of the capital share parameter in the modern production function. *Main finding:* The differences in relative efficiency required to generate delays in starting dates of the lengths observed in the historical data are reasonable for all capital shares above 0.40.

We then compute the entire equilibrium path of these late starters assuming that their efficiency levels relative to the leader never change. The main finding is that the gap in incomes between late and early starters never narrows. Large differences in incomes exist even after the late starters are in the modern economic growth phase. In fact, the gap between the leader and late starters increases for some time after the laggards have started

the transition to modern economic growth. This is the case even though the transition period of late starters is shorter compared to early starters. The difference in relative efficiencies between late and early starters needed to generate a given factor difference in per capita outputs when both sets of countries are experiencing modern economic growth again depends upon the capital share parameter.

The final set of experiments allows for a one-time increase in a country's relative efficiency parameter. We assume that the change is unexpected from the standpoint of the late starter and viewed as permanent in nature. We then compute the equilibrium path relative to the leader's level and determine the country's output relative to the leader subsequent to the change. *We find that the late starter's path of output relative to the leader subsequent to the change in its efficiency parameter is consistent with the experience of growth miracle countries such as Japan, but only if the capital share is between one-half and two-thirds.*

The finding that capital's share must be large for the unified theory to be a successful theory of the evolution of international income levels has important implications for the size of investment in intangible capital. Namely, it implies that the size of this investment is a large fraction of GDP. Investment in intangible capital goes unmeasured in the national income and product accounts. Thus, it is not possible to determine whether a large capital share is plausible by examining national account data. One must examine micro evidence to determine the plausibility of a large capital share. Thus, we conclude this section by examining the micro evidence on the size of unmeasured investment in the economy. We conclude from this evidence that the size of

unmeasured investment in the economy is as large as the size predicted by the unified theory.

Delays in Starting Dates

We first examine whether the unified theory predicts large delays in the start of the transition to modern economic growth that some countries have experienced. In particular, we determine the size of the difference in efficiency required to delay the start of the transition to modern economic growth by a certain number of years.

For the purpose at hand, it is important to provide a more thorough picture of the different starting dates for the transition corresponding to the experiences of individual countries. An issue is how to date the start of modern economic growth. Our definition of the *start* of modern economic growth is the earliest point in a country's history with the property that the trend growth rate is 1 percent or more for all subsequent time.[20] Figure 7 shows the path of output in a number of countries relative to the industrial leader going back to 1800. As can be seen, starting dates vary substantially across countries. Mexico started the transition to modern economic growth sometime between 1800 and 1850; Japan started sometime between 1850 and 1900. Brazil started in the early twentieth century, and India started its transition sometime between 1950 and 1980. As a result of these different starting dates, the disparity in income has increased.

The key expression for determining the delay in the starting date associated with differences is equation (2.12), which rewritten in relative efficiencies is

$$E_S^i A_{St} < \left(\frac{r_{Mt}}{\theta}\right)^{\theta} \left(\frac{w_{Mt}}{1-\theta}\right)^{1-\theta}. \tag{4.1}$$

[20] The concept of trend employed here is a highly smoothed path of per capita income.

A country will not use the modern production function as long as the relation given by (4.1) is satisfied. Once a country's efficiency, $E_S A_{St}$, exceeds the critical level given by the right-hand side of (4.1), which it must, the country begins its transition to modern economic growth. Assuming as we do that relative efficiencies associated with the traditional production function do not differ across countries, the rental prices of land and labor will not differ much across countries over the periods when each country specializes in the traditional production function.[21] Consequently, this critical level of efficiency will not differ much across countries. It follows that the difference in starting dates between two countries *i* and *j*, with different relative efficiencies, is approximately given by the dates t_i and t_j for which

$$E_S^i A_{St_i} = \left(\frac{r_M}{\theta}\right)^{\theta} \left(\frac{w_M}{1-\theta}\right)^{1-\theta} = E_S^j A_{St_j} . \tag{4.2}$$

It is not obvious looking at equation (4.2), but the required relative efficiency E_S^i / E_S^j that gives rise to a particular delay in the start of the transition depends on the size of the capital share in the modern production function. The reason for this is that the required factor difference in relative efficiencies equals the factor difference in the stock of pure knowledge, A_s, between starting dates. It follows that the required relative efficiency difference is smaller for larger increases in the stock of pure knowledge between starting dates. The size of the increase in the stock of pure knowledge depends importantly on its asymptotic growth rate, γ_S. The value of this parameter is calibrated so that the growth rate of per capita output associated with the steady-state rate of the

[21] They are roughly equal because the rental prices will not be constant in all periods that the economy specializes in the traditional production function. This is because agents will start to accumulate more capital per household member in anticipation of the modern production function being used.

modern-growth-only model, given by $(1 + \gamma_S)^{1/(1-\theta)}$, equals 2 percent per year. Thus, the calibrated value of γ_S and hence the size of the increase in pure knowledge between starting dates t_j and t_i, depends on capital's share in the modern growth production function.

We now compute the efficiency of the early starter relative to a late starter required to generate a given delay in the transition to modern economic growth. We do this for a range of the capital share parameters, since the value of capital's share is not well restricted. For each capital share value, we recalibrate the asymptotic growth rate of pure knowledge, γ_S, and the value of A_{s0} so that the country with $E_s = 1$ always starts its transition in 1700. These are the only parameters whose values are changed in the experiments.

We assume that late starters are endowed with an initial capital stock equal to the steady-state level associated with the classical model of subsection 2A. For the purpose of determining the date at which an economy starts to use the modern growth function, it is not necessary that we fully specify the population growth function of the late starters. In particular, it is not necessary to specify the population growth function for consumption levels sufficiently greater than the constant consumption level, c_m, associated with the pre-1700 period. For consumption levels below this, we use a population growth function with a sufficiently large and positive slope at c_M and for which $g(c_M) = (1+\gamma_M)^{1/(1-\phi-\mu)}$. These assumptions ensure that the living standard in a late starter is roughly constant prior to the period it begins its transition.

Table 2 reports the efficiency of the early starter relative to the late starter required to generate a 100-year, a 200-year, and a 250-year delay in the transition to

modern economic growth. These delays roughly represent the difference in the start of the transition to modern economic growth between England and Mexico, England and Japan, and England and India. As Table 2 shows, the factor difference in efficiency needed for a given delay decreases as the modern production capital share increases. The size of the required difference needed to delay the start of development for 250 years is plausible for all values of θ in Table 2, with $\theta = 0.40$ probably at the lower bound of plausible values.

Table 2. Required Factor Difference in Relative Efficiencies for Delays

θ	1800 Start	1900 Start	1950 Start
.40	1.60	3.2	5.7
.50	1.25	2.5	4.0
.60	1.20	2.2	3.3
.70	1.18	1.9	2.5

No Catch-Up After the Transition

A number of countries, many of which are located in Latin America, started their transitions to modern economic growth in the nineteenth century. Despite this, these countries have failed to eliminate the gap with the leader over the last century. We now examine whether the model can account for this feature of the data. In particular, we seek to determine if the model predicts a narrowing of income levels once a country begins modern economic growth absent any changes in relative efficiency.

We address this question by examining whether the model absent any assumed subsequent changes in relative efficiencies predicts a narrowing or widening of income levels between early and late starters. In particular, we now compute the equilibrium paths of per capita output for the model economies associated with the required

differences in relative efficiencies reported in Table 2. We also report their relative incomes.

Before undertaking these experiments, it is necessary to address two issues. First, it is necessary to specify the population growth rate function for the late starters in these experiments because increases in population affect the size of the increases in per capita output over the transition. For this specification, we simply use the post-1800 population growth rates of Mexico for the model economy that starts its transition in 1800, the post-1900 population growth rates of Japan for the model economy that starts its transition in 1900, and the post-1950 population growth rates of India for the model economy that starts its transition in 1950. These population growth data are taken from Lucas (2002, Table 5.1). Second, for capital share values that reflect a broad concept of capital, it is necessary to adjust output by the amount of investment in intangible capital. This adjustment must be made in order to compare the predictions of the model with the national income and product account data, because the latter fails to measure investments in intangible capital.

A country's unmeasured investment as a fraction of its measured output can be determined given the decomposition of the capital share between its physical capital and intangible capital components. For a given total capital share, the physical capital component can be calibrated to the ratio of investment to physical capital to measured GDP in the leader countries of roughly 20 percent. In particular, the share parameters can be calibrated to the steady state of the modern growth-only-economy using this observation from the leader countries. With value of the individual share parameters in

the modern growth production function, it is possible to compute the amount of unmeasured investment at any date of the equilibrium path.

Table 3 reports the size of the intangible capital share parameter and the asymptotic ratio of intangible capital investment to GDP for each of the total capital share values considered in Table 2. As the total capital share increases, both the intangible capital share and the intangible capital investment share of GDP increase. The sizes of the unmeasured investment shares range from 0.0 for $\theta = 0.40$ to 0.50 for $\theta = 0.70$.

Table 3. Implied Intangible Capital Share and Investments

θ	θ_I	$X_I/(Y-X_I)$
.40	.00	.00
.50	.28	.26
.60	.41	.41
.70	.53	.62

Figure 8 plots the path of per capita GDP for late starters relative to the leader over the 1700 to 2050 period. The paths correspond to the case where $\theta = 0.40$. The paths are essentially the same for the other capital share values. For this reason, we do not report their paths in the paper. Asymptotically, the model is just the steady state of the modern growth model of subsection 2B, and so income differences are just $(E_s^i / E_s^j)^{1/(1-\theta)}$. For the 1800 starter, the asymptotic relative income level is 50 percent of the leader, for the 1900 starter it is 16 percent, and for the 1950 starter it is 6 percent.

Most of the difference in relative incomes in 2000 is the consequence of the poor country starting the development process later. However, even after starting to develop, a late starter's disparity with the leader increases, although at a much slower rate than before. There are two reasons for this. First, the disparity continues to increase because

the traditional production function is still widely used at the start of the transition and the growth rate of TFP associated with the traditional production function is lower than the growth rate of TFP associated with the modern production function. Second, the population growth in these countries tends to be higher compared to the leader over the comparable period. The disparity with the leader stops increasing only after the modern production function starts being used on a large scale. For the 1800 starter, the disparity stops increasing around 1900. For the 1900 starter, the disparity stops increasing around 2000. And for the 1950 starter, the disparity stops increasing around 2050.[22] The increase in disparity over the 1950–2000 period for the 1950 starter is consistent with the fact that many sub-Sahara African countries have fallen further behind the leader in the 1950–2000 period despite experiencing absolute increases in living standards over this period.

Laggards do experience larger increases in their income over their transition periods compared to earlier starters. For example, the country that starts its transition in 1700 realizes a factor increase of 1.2 in its per capita income by 1750. In comparison, the country that starts its transition in 1900 realizes a factor increase of 2 in its per capita income over the next 50 years.[23] The reason for this difference is that the growth rate of knowledge associated with the modern production function is initially low, but rises over time. Thus, TFP growth in the modern production function over a late starter's transition period is higher compared to an earlier starter's transition period. This gives late starters an inherent advantage.

[22] This is a key difference between our formulation and that of Ngai (2000). Ngai examines the effect of policy on the starting date within Hansen and Prescott's overlapping generations model. In contrast, she finds that some part of the income gap will be eliminated once poor countries start their transitions.

[23] This assumes the same population growth functions for both economies.

The data needed to verify whether this pattern exists are not readily available. In particular, per capita output numbers going back to the eighteenth century exist for only a limited number of countries. Although it is not possible to say whether transition periods have become shorter over time, there is strong evidence that late starters have been able to double their incomes in far shorter time periods compared to earlier starters.

Figure 9 documents this general pattern. It plots the number of years a country took to go from 10 percent to 20 percent of the 1985 U.S. per capita income level versus the first year that country achieved the 10 percent level. The 1985 U.S. level was 20,000 in 1990 dollars. The set of countries considered had at least 1 million people in 1970 and had achieved and sustained per capita income of at least 10 percent of the 1985 U.S. level by 1965. There are 56 countries that fit these criteria and for which data are available. Of these 56 countries, all but four managed to double their per capita income by 1992. The four exceptions all had protracted armed insurgencies that disrupted their development.

The difference in the length of the doubling period between the sets of late and early starters is dramatic. For early starters, which are those achieving 10 percent of the 1985 U.S. level before 1950, the median length of the doubling period is 45 years. For late starters, defined as those achieving 10 percent of the 1985 U.S. level after 1950, the median length of the doubling period is 15 years. The choice of starting level is not important. A similar pattern emerges when the starting level is fixed at 5 percent and at 20 percent of the 1985 U.S. level.

Although the model absent changes in relative efficiency infers an advantage to late starters, quantitatively it is inconsistent with the number of years in which many late starters have been able to double their income. Many late starters that doubled their

income in less than a 35-year period after 1950 did in fact narrow the gap with the leader over that period. The unified theory absent changes in relative efficiencies does not predict any catch-up for late starters. For the theory to account for this catch-up, it must consider changes in relative efficiency in a given country over time.

Catch-Up and Growth Miracles

We now examine whether the theory can account for the record of catch-up. A key feature of the evolution of international income levels is that many countries have been able to narrow the gap with the leaders, with some realizing large increases in output relative to the leader in a relatively short period of time. Countries such as Botswana, China, Japan, South Korea, and Taiwan were all able to double their living standards in less than a decade at some point in time over the post-1950 period. These growth miracles are a relatively recent phenomenon and are limited to countries that were relatively poor prior to undergoing their miracle. No country at the top of the income distribution has increased its per capita income by a factor of 4 in 25 years, and the leader has always taken at least 80 years to quadruple its income.

To account for the catch-up, including growth miracles, the theory, therefore, requires an increase in the efficiency of a country relative to the leader.[24] In light of the Parente and Prescott (2000) theory, these changes in relative efficiency are easy to understand. Namely, they reflect policy changes. Following an improvement in policy that leads to a significant and persistent increase in efficiency, the theory predicts that the income of a late starter will go from its currently low level relative to the leader to a

[24] Additionally, an increase in efficiency can hasten the start of the transition to modern growth for countries that have not already begun this phase of development.

much higher level. As it does, its growth rate will exceed the rate of modern growth experienced by the leader countries, and the gap in incomes will be narrowed.

We now consider an increase in a late starter's relative efficiency. In particular, we examine whether the unified theory can account for the growth miracle of Japan.[25] Figure 10 depicts the path of per capita output for the Japanese and U.S. economies over the 1900 to 1995 period. There is really nothing special about Japan versus other economies that similarly experienced growth miracles. The precise time period of the Japanese growth miracle we consider in the analysis is the 1957–69 period. We choose this period because by 1957 Japan had fully recovered from the wartime disruptions. Moreover, this period is one of the most dramatic in terms of Japan's catch-up. In this 12-year period, per capita GDP doubled from 25 percent of the leader to 50 percent of the leader (Summers and Heston, 1991). This catching up was not the result of the leaders' growth rate slowing down. Indeed, U.S. per capita GDP grew by 40 percent in this period. The Japanese economy in this period is a dramatic example of catching up.

In the experiment, we assume that there is an unexpected increase in 1957 in the relative efficiency of the model economy, which started its transition in 1900, to the leader's level. This assumption is made because the data suggest that Japan in the 1957–69 period was converging to the U.S. balanced growth path. In calculating the equilibrium path of the model economy following this increase, we take the initial population to be the population corresponding to the equilibrium path of the model economy that starts the transition in 1900. The initial capital stock is assumed to be such

[25] Ngai (2000) studies this same issue within the Hansen and Prescott model.

that per capita GDP relative to the leader equals 25 percent.[26] The population growth rate function for the model economy is the same as before and is based on Japanese population dynamics.

The important finding is that the total capital share must be large for an economy to take 12 years to move from 25 percent to 50 percent of the leader. Figure 11 plots the path of per capita GDP predicted by the model economy over this period for various values of θ. For a value of θ equal to 0.40, the predicted path shows too large an increase over the period. At the other end of the range, namely, $\theta = 0.70$, the predicted path shows too small an increase over this time period. This leads us to conclude that capital share values in the range of 0.55 and 0.65 are consistent with the growth miracles. [27]

It is possible to introduce this increase in efficiency in the poor country at a much earlier date, say in 1800. The theory does not, however, predict that the poor country will experience a growth miracle. The theory, therefore, is consistent with the fact that growth miracles are a relatively recent phenomenon. Growth miracles are a relatively recent phenomenon because, as Figure 8 shows, differences in relative incomes between the low-efficiency and high-efficiency countries widen over time before leveling off. This widening is due to growth in the stock of pure knowledge associated with the modern production function, which the high-efficiency country uses from a very early date. Thus, as one goes back in time, the gap that a low-efficiency country could close by becoming a high-efficiency country becomes smaller and smaller. Obviously, if the gap is less than

[26] In the case where capital is broadly defined, we assume the initial mix of physical and intangible capital is optimal in the sense that returns would be equal.

[27] There are a number of reasons to believe that capital's share may be somewhat less than 0.60. For one, we abstracted from leisure. For another, we abstracted from household durables. For an in-depth discussion of this issue, see Parente and Prescott (2000).

50 percent, the low-efficiency country can never double its income in less than a decade. For the same reason, the unified theory is consistent with the fact that late starters have been able to double their incomes in far shorter times compared to early starters.

The theory is also consistent with the fact that growth miracles are limited to countries that were initially poor at the time their miracles began. Growth miracles are limited to this set of countries because a growth miracle in the theory requires a large increase in a country's relative efficiency. A large increase in efficiency can only occur in a poor country with a currently low efficiency parameter. This rules out a rich country, which by definition uses its resources efficiently.

Unmeasured Investment

For capital shares that are consistent with the evolution of international income levels, the implied size of unmeasured investment is between 35 and 55 percent of GDP. Are these intangible capital investment share numbers plausible? This is not an easy question to answer. The difficulty in coming up with measures of the size of intangible capital investment is that the national income and product accounts (NIPA) treat investments in intangible capital as ordinary business expenses. Parente and Prescott (2000) attempt to estimate the size of intangible capital investment in the U.S. economy. They conclude that the size of this investment may be as large as 50 percent of GDP. In constructing their estimates, Parente and Prescott (2000) use the principle implied by theory that *investment* is any allocation of resources that is designed to increase future production possibilities. Using this principle, they identify such activities as starting up a new business, learning-on-the-job, training, education, research and development, and

some forms of advertising as investments in intangible capital.[28] Such estimates are consistent with capital share values between one-half and two-thirds.

5. Catching Up

The implication of the theory is that countries will be rich if they do not constrain production units as to which technologies can be operated and the manner in which a given technology can be operated. Currently poor countries will catch up to the industrial leaders in terms of production efficiency if existing barriers to efficient production are eliminated and an arrangement is set up to ensure that barriers will not be re-erected in the future. The removal of such constraints is a necessary condition for catching up. As discussed in Section 3, there is strong evidence that suggests that these constraints exist to protect the interests of industry groups vested in the current production process. As such, their removal is likely to be contentious. For this reason, it is instructive to examine the record on catch-up in greater depth for the purpose of determining the reasons for circumstances under which barriers to efficient use of technology were reduced and catching up with the efficiency leader occurred.

Catching up is not uniform across regions in this period, as can be seen in Figure 12. Latin America began modern economic growth in the late nineteenth century and has not subsequently closed the living standards gap with the industrial leader. Its per capita income remained at roughly 25 percent of the industrial leader throughout the twentieth century. In comparison, Asian countries with the exception of Japan began modern

[28] Additionally, McGrattan and Prescott (2002) estimate the size of unmeasured investment in the corporate sector only and conclude that it is roughly 10 percent of GDP.

economic growth later. This set of countries experienced significant catching up in the 1970–2000 period.

The large Western European countries, namely, Germany, Italy, and France, caught up to the industrial leader in the post–World War II period after trailing the leader for 100 years. Modern economic growth in these countries began about 1840. At that time, their living standard was about 60 percent of the industrial leader, which at that time was the United Kingdom. For nearly 100 years, these countries maintained an income level that was about 60 percent that of the industrial leader. In the post–World War II period, output per hour worked in these countries, which is a good measure of living standards because it recognizes the value of nonmarket time, increased from 38 percent of the U.S. level in 1950 to 73 percent in 1973 and to 94 percent in 1992. Today, most of the difference in per capita output between the Western European countries and the United States is accounted for in differences in the fraction of time that people work in the market, and not in the efficiency with which resources are used.

Another important example of catching up is the U.S. development experience in the 1865–1929 period. In 1870, U.K. per capita GDP was nearly a third higher than that of the United States. By 1929, the United Kingdom's per capita GDP was a third lower than that of the United States. The dramatic growth performance of the United States in this period is an important fact that needs to be explained.

Reasons for Catching Up

We begin with the Asian catching-up observation. Countries such as South Korea, Taiwan, and Japan were forced to adopt policies that did not block efficient production as a condition for support from the United States. Further, the need to finance national

defense made protecting those with vested interests in inefficient production too expensive to South Korea and Taiwan. These development miracles along with the Hong Kong and Singapore growth miracles made it clear to the people of the democratic states in the region that the policy that their elected representatives followed mattered for their living standard. Their elected representatives had no choice but to cut back on protecting industry insiders with vested interests in inefficient production or be voted out of office.

The rapid development of China began in 1978 when the Chinese government became more decentralized, with much of the centralized planning system dismantled. Although the central government gave more power to regional governments, it did not give the regional governments the right to restrict the flow of goods across regions. In fact, when individual regions attempted to erect trade barriers in the late 1980s and early 1990s, the central government immediately took steps to restore the free flow of goods and services.[29] The resulting competition between businesses in different provinces led to rapid growth in living standards.

The comparison of Russia's performance under capitalism with China's is interesting and informative. Russia's experiment with capitalism to date can only be considered a failure, as its output has actually contracted since 1992. In contrast to China, there is no free trade club in Russia. Migration of individuals between regions is restricted, and local and regional governments have the power to discriminate against producers from other member states operating within their borders. Parente and Riós-Rull (2001) argue that establishing a decentralized system with competition between regions in Russia was undoubtedly a much more difficult endeavor compared to China for a

[29] See Young (forthcoming).

number of reasons. First, by being more industrialized at the time of its transition, Russia had more vested interest groups. Second, Soviet central planners concentrated industry in particular regions, without an economic justification for such locations.

Turning now to the questions of why the United States caught up with and surged past the United Kingdom in the 1865–1929 period and why Western Europe caught up with the United states in terms of labor productivity in the 1957–93 period, our answers are as follows. The answer to the first question is that the United Sates was and continues to be a free trade club, while the United Kingdom was not a member of a free trade club in this earlier period. Our definition of a *free trade club* is as follows. A set of states constitutes a free trade club if it meets two conditions. Member states cannot impose tariffs and other restrictions on the import of goods and services from other member states. In addition, member states must have a considerable degree of economic sovereignty from the collective entity. Just as no single state is able to block the movement of goods between states, the collective entity cannot block the adoption of a superior technology in one of its member states. Thus, a free trade club in our definition is far more than a set of countries with a free trade agreement.

In democratic states with legislatures representing districts, vested interests in other districts have a limited ability to block the adoption of technology in a given district if the citizens of the given district want that technology adopted. In the United States, for example, Toyota was able to locate an automobile plant with its just-in-time production in Tennessee in 1985. Those with vested interests in the less efficient technology in Michigan and other states with a large automotive industry were not able to prevent this from happening. The people in Tennessee wanted the large construction project in their

state and the high paying jobs in the automobile factory. Thus, the United States is a free trade club. With the formation of NAFTA and the recent approval of the free trade agreements with Chile and Singapore, the set of states constituting the free trade club to which the U.S. states belong may be getting larger.

The European Union has become an equally important free trade club. Its states enjoy even greater sovereignty than do U.S. member states. However, the German state cannot block the Toyota introduction of just-in-time production in Wales even though German politicians would if they could in response to domestic political pressure. If Toyota starts gaining market share, it will not be long before the auto industry throughout Europe adopts the superior technology, and productivity in the production of automobiles increases. This is just competition at work.

The historical statistics lend strong empirical support to the theory that a trading club arrangement results in greater efficiency of production. Table 4 reports labor productivity for the original members of what became the European Union and the labor productivity of members that joined in the 1970s and 1980s. Productivities are reported for an extended period before the EU was formed as well as for the period subsequent to its creation.

The Treaty of Rome was signed in 1957 by Belgium, France, Italy, Netherlands, Luxembourg, and West Germany to form the union. In 1973 Denmark, Ireland, and the United Kingdom joined. In 1981 Greece joined, followed by Portugal and Spain in 1986. The most recent additions are Austria, Finland, and Sweden in 1995.

One striking fact is that prior to forming the European Union, the original members had labor productivity that was only half that of the United States. This state of

affairs persisted for over 60 years with no catching up. However, in the 36 years after forming what became the EU, the Treaty of Rome signers caught up with the United States in terms of labor productivity. The factor leading to this catch-up is an increase in the efficiency with which resources are used in production. Changes in capital/output ratios are of little significance in accounting for the change in labor productivity.

Also reported in Table 5 is the productivity of the EU countries that joined the union in 1973. These countries experienced significant productivity catch-up subsequent to joining the union. It will be interesting to see if Greece and Portugal, the two EU countries that have far lower productivity than the other EU members, continue to improve their relative productivity performance.

Another interesting comparison is between the productivity performance of Switzerland and the Western European countries that did not join the EU until 1995. Norway was not included in this set of countries because of the large size of its oil industry. We label this set of four countries *other*. Table 5 reports labor productivities of these other countries relative to the original EU countries.

The important finding is that the original EU countries and the other countries are equally productive in the prewar period. In the 36 years from 1957 to 1993, the other countries fell from 1.06 times as productive as the original EU countries to only 0.81 as productive in 1993. This constitutes strong empirical evidence that membership in the EU fosters higher productivity.

Table 4. Labor Productivities of European Union Members as a Percentage of U.S. Productivity[a]

Year	Original Members	Members Joining in 1973
1870	62	
1913	53	
1929	52	
1938	57	
1957	53	57
1973	78	66
1983	94	76
1993	102	83
2002	101	85

[a] The prewar numbers are population weighted labor productivity numbers from Maddison (1995). The postwar numbers are also population weighted and were obtained from Maddison's Web page, http://www.eco.rug.nl/GGDC/index-series.html#top.

Table 5. Labor Productivity of Other Western European Countries as a Percentage of Original EU Members[a]

Year	Others / Original
1900	103
1913	99
1938	103
1957	106
1973	96
1983	85
1993	81

[a] The prewar figures are from Maddison (1995). For this period, GDP per capita is used as a proxy for productivity. The postwar numbers are also population weighted and were obtain from Maddison's Web page, http://www.eco.rug.nl/GGDC/index-series.html#top.

A free trade club, which prohibits individual states from discriminating against the goods produced in other member states and against producers from other member states operating within their borders, has the advantage that industry insiders in the various member states face elastic demand for what they supply. As a consequence, they are not hurt by the adoption of more efficient production methods as the increase in output leads to an increase in employment in that industry. If demand were inelastic, an increase in efficiency would lead to a fall in employment, something which industry insiders strongly oppose.

Industry studies document the effect of free trade of goods and services on the adoption of better technology and work practices. Galdon and Schmitz (1998), for example, document the effect of increased competition in iron ore mining in the 1980s.

Increased competition from Brazilian iron ore mines had major consequences for productivity in U.S. mines. Output per unit of input increased by a factor of 2 as competition made it in the interest of specialized factor suppliers to permit the doubling of productivity. Ferreira and Rossi (forthcoming) document large increases in output per worker in 16 industries in Brazil at the two-digit level following the trade liberalization in the early 1990s. After declining at an annual rate of 1.6 percent per year from 1985 to 1990, it increased at a rate of 6 percent per year thereafter. The increases in productivity were associated with a decline in employment and hours.

We turn now to Latin America and why Latin America failed to catch up. There was no free movement of goods and people between the set of relatively sovereign states. A consequence of this is that often industry insiders in the sovereign states faced inelastic demand for their products or services, and this led them to block the adoption of more efficient production practices. If Brazil were to decentralize and restrict the authority of its central government to be like the United States in the 1865–1930 period, Brazil would quickly become as rich as Western Europe and the United States, or maybe richer.

6. Concluding Remarks

Will the whole world be rich by the end of the twenty-first century? The implication of the theory reviewed in this chapter is that a country will catch up to the leading industrial countries only if it eliminates the constraints relating to the use of technology. Although it is clear what a country must do to become rich, it is not clear whether a country will have either the political will or political power to make the necessary reforms. Removal of the constraints to the efficient use of resources is bound to be contentious, because such constraints typically exist to protect specialized groups of

factor suppliers and corporate interests. As recent events in Argentina show, these groups can overthrow a government.

The increase in the number of free trade clubs in the last decade, the central Andean Community and North America, for example, is evidence that some countries have achieved the political will to reduce these constraints. However, the lack of the emergence of free trade clubs in many other regions of the world, particularly Africa, the Indian subcontinent, and South America, is evidence of a lack of political will. A thorough understanding of why one country has this political will and another does not is something we are currently lacking. If we had this understanding it might be possible to determine what should be done to minimize the resistance to reform by groups with interests vested to current production processes.

A first step in addressing this issue is to understand how constraints to the efficient use of resources come to exist in the first place. Surely, many constraints exist to protect the vested interests of individuals in the status quo. What we really seek to understand is the mechanism by which these groups and their interests succeed in getting these constraints put in place. Policy, namely, the imposition of constraints on the efficient use of resources, is undoubtedly the outcome of a game between policymakers and the economy's actors. Consequently, fruitful research in this area will most likely require a game-theoretical approach. Some progress is being made in this area. Grossman and Helpman (1994), Krusell and Ríos-Rull (1996; 2002), Holmes and Schmitz (1995; 2001), McDermott (1999), Kocherlakota (2000), Ngai (2000), Bridgman, Livshits, and MacGee (2001), Parente and Ríos-Rull (2001), Samaniego (2001), Teixeira (2001), and

Parente and Zhao (2002) all deal with this issue. In our view, this area of research will dominate the study of development and growth in the years to come.

References

Bairoch, P. *Economics and World History: Myths and Paradoxes* (Chicago: University of Chicago Press, 1993).

Becker, G. S. "An Economic Analysis of Fertility." In Richard Easterlin, ed., *Demographic and Economic Change in Developed Countries*. Universities-National Bureau Conference Series, no. 11 (Princeton, N.J.: Princeton University Press, pp. 209–40, 1960).

Becker, G. S., K. M. Murphy, and R. Tamura. "Human Capital, Fertility, and Economic Growth." *Journal of Political Economy* 98 (1990), S12–37.

Bridgman, B., I. Livshits, and J. MacGee. "For Sale: Barriers to Riches." Unpublished Manuscript, Federal Reserve Bank of Minneapolis, 2001.

Djankov, S., R. La Porta, F. Lopez-de-Silanes, and A. Shleifer. "The Regulation of Entry." *Quarterly Journal of Economics* 117 (2002), 1–37.

Doepke, M. "Fertility, Income Distribution, and Growth." Doctoral Dissertation, University of Chicago, 2000.

Ferreira-Cavalcanti, P., and J. L. Rossi. "Trade Barriers and Productivity Growth: Cross-Industry Evidence." *International Economic Review* (forthcoming).

Galdon-Sanchez, J. E., and J. A. Schmitz, Jr. "Tough Markets and Labor Productivity: World Iron-Ore Markets in the 1980s." *American Economic Review* 92 (2002), 1222–35.

Galor, O. and D. Weil. "Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and Beyond." *American Economic Review* 90: (2000), 806–28.

Gollin, D. “Getting Income Shares Right.” *Journal of Political Economy* 110 (2002), 458–74.

Grossman, G. M., and E. Helpman. “Protection for Sale.” *American Economic Review* 84 (1994), 833–50.

Hall, R. E., and C. I. Jones. “Why Do Some Countries Produce So Much More Output Per Worker than Others?” *Quarterly Journal of Economics* 114 (1999), 83–116.

Hansen, G. D., and E. C. Prescott. “Malthus to Solow.” *American Economic Review* 92 (2002), 1205–17.

Hendricks, L. “How Important Is Human Capital for Development? Evidence from Immigrant Earnings.” *American Economic Review* 92 (2002), 198–219.

Hendricks, L. “Taxation and the Intergenerational Transmission of Human Capital.” *Journal of Economic Dynamics and Control* (forthcoming).

Herrendorf, B., and A. Teixeira. “Monopoly Rights Can Reduce Income Big Time.” University of Carlos III working paper, July 2003.

Holmes, T. J., and J. A. Schmitz, Jr. “Resistance to New Technology and Trade Between Areas.” *Federal Reserve Bank of Minneapolis Quarterly Review* 19 (Winter 1995), 2–17.

Holmes, T. J., and J. A. Schmitz, Jr. “A Gain from Trade: From Unproductive to Productive Entrepreneurship.” *Journal of Monetary Economics* 47 (2001), 417–46.

Hornstein, A. and E. C. Prescott. “The Firm and the Plant in General Equilibrium Theory.” In R. Becker, M. Boldrin, R. Jones, and W. Thomson, eds., *General*

Equilibrium, Growth, and Trade. Vol. 2. The Legacy of Lionel McKenzie (San Diego: Academic Press, pp. 393–410, 1993).

Johnson, D. G. "Population, Food, and Knowledge." *American Economic Review* 90 (2000), 1–14.

Klenow, P., and A. Rodriguez-Clare. "Economic Growth: A Review Essay." *Journal of Monetary Economics* 40 (1997), 597–617.

Kocherlakota, N. R. "Building Blocks for Barriers to Riches." Unpublished Manuscript, Federal Reserve Bank of Minneapolis, 2000.

Krusell P., and J.-V. Ríos-Rull. "Vested Interests in a Positive Theory of Stagnation and Growth." *Review of Economic Studies* 63 (1996), 301–29.

Krusell, P., and J.-V. Ríos-Rull. "Politico-Economic Transition." *Review of Economic Design* 7 (2002), 309–29.

Kuznets, S. *Modern Economic Growth* (New Haven, Conn.: Yale University Press, 1966).

Lucas, R. E., Jr. *Lectures on Economic Growth* (Cambridge: Harvard University Press, 2002).

Maddison, A. *Dynamic Forces in Capitalist Development: A Long-Run Comparative View* (Oxford University Press, 1991).

Maddison, A. *Monitoring the World Economy: 1820–1992* (Paris: Organisation for Economic Co-operation and Development, 1995).

Malthus, T. R. "First Essays on Population" [1797]. *Reprints of Economic Classics* (New York: Augustus Kelley, 1965).

McDermott, J. “Mercantilism and Modern Growth.” *Journal of Economic Growth* 4 (1999), 55–80.

McGrattan, E. R., and E. C. Prescott. “Taxes, Regulations, and the Value of U.S. Corporations: A General Equilibrium Analysis.” Staff Report 309, Federal Reserve Bank of Minneapolis, August 2002.

Mokyr, J. *The Lever of Riches: Technological Creativity and Economic Progress* (New York: Oxford University Press, 1990).

Ngai, L. R. “Barriers and the Transition to Modern Economic Growth.” Unpublished Manuscript, University of Pennsylvania, 2000.

Parente, S. L., and E. C. Prescott. “Monopoly Rights: A Barrier to Riches,” *American Economic Review* 89 (1999), 1216–33.

Parente, S. L., and E. C. Prescott. *Barriers to Riches* (Cambridge: MIT Press, 2000).

Parente, S. L., and J.-V. Ríos-Rull. “The Success and Failure of Economic Reforms in Transition Economies.” Unpublished Manuscript, University of Illinois, 2001.

Parente, S. L., and R. Zhao. “From Bad Institution to Worse: The Role of History in Development.” Unpublished Manuscript, University of Illinois, 2002.

Ricardo, D. “On the Principals of Political Economy and Taxation” [1817]. In Pier Sraffa, ed., *The Works and Correspondences of David Ricardo*, Vol. 1 (Cambridge: Cambridge University Press, 1951).

Samaniego, R. M. “Does Employment Protection Inhibit Technology Diffusion?” Unpublished Manuscript, University of Pennsylvania, 2001.

Schmitz, J. A., Jr. “Government Production of Investment Goods and Aggregate Labor-Productivity.” *Journal of Monetary Economics* 47 (2001), 163–87.

Solow, R. M. *Growth Theory: An Exposition* (New York: Oxford University Press, 1970).

Summers, R. and A. Heston. "The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950–1988." *Quarterly Journal of Economics* 106 (1991), 327–68.

Tamura, R. "Fertility, Human Capital, and the 'Wealth of Nations.'" Doctoral Dissertation, University of Chicago, 1988.

Teixeira, A. "Effects of Trade Policy on Technology Adoption and Investment." Unpublished Manuscript, 2001.

Young, A. "The Razor's Edge: Distortions and Incremental Reform in the People's Republic of China." *Quarterly Journal of Economics* (forthcoming).

Figure 1: Evolution of International Incomes: 1700–1990
(Fraction of Leader)

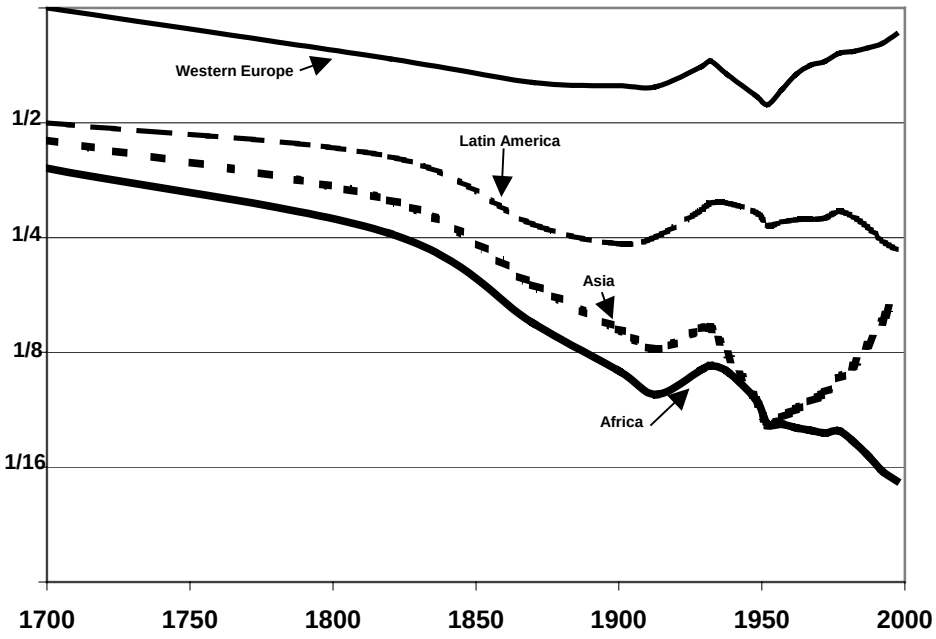


Figure 2: The Leader's Per Capita GDP Relative to Pre-1800 Level

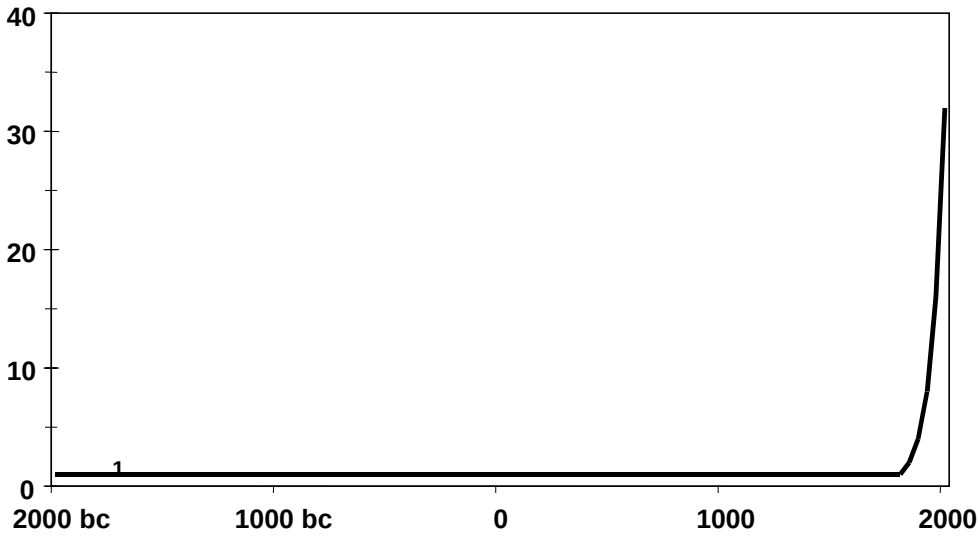


Figure 3: Population Growth Function $g(c)$

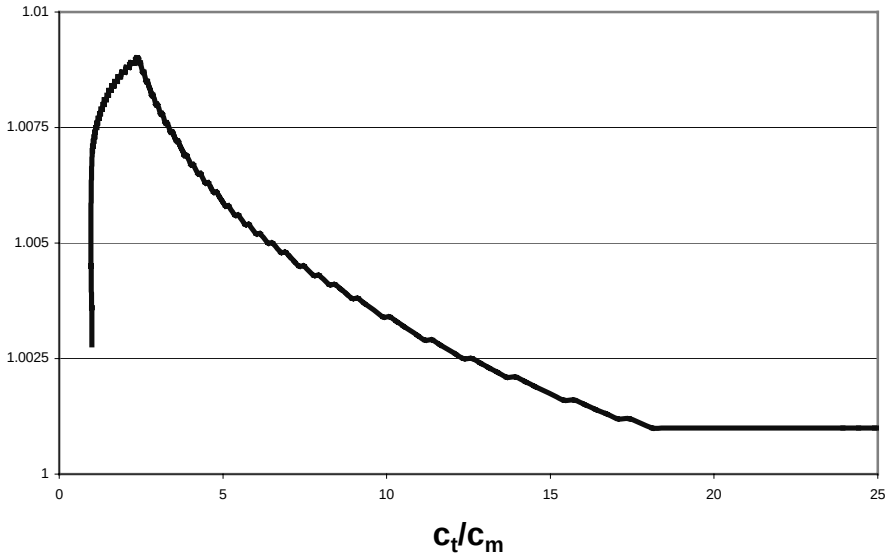


Figure 4: Per Capita Output Relative to 1700

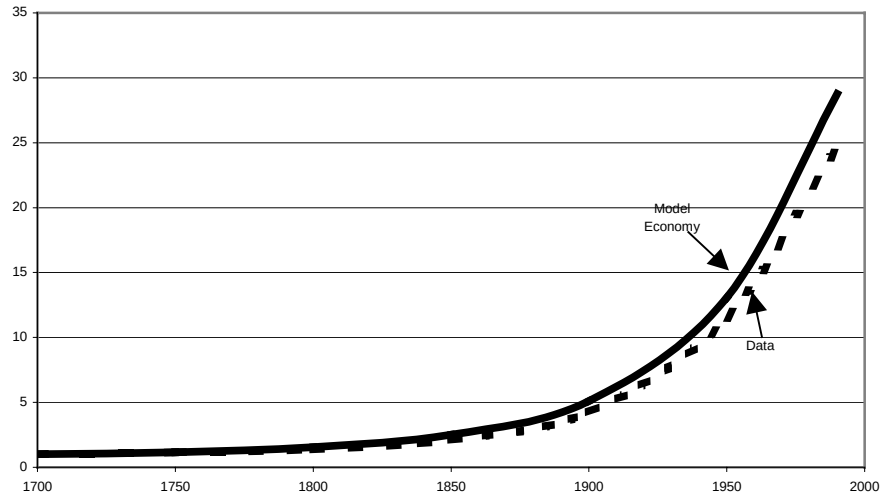


Figure 5: Growth Rate of per Capita Output

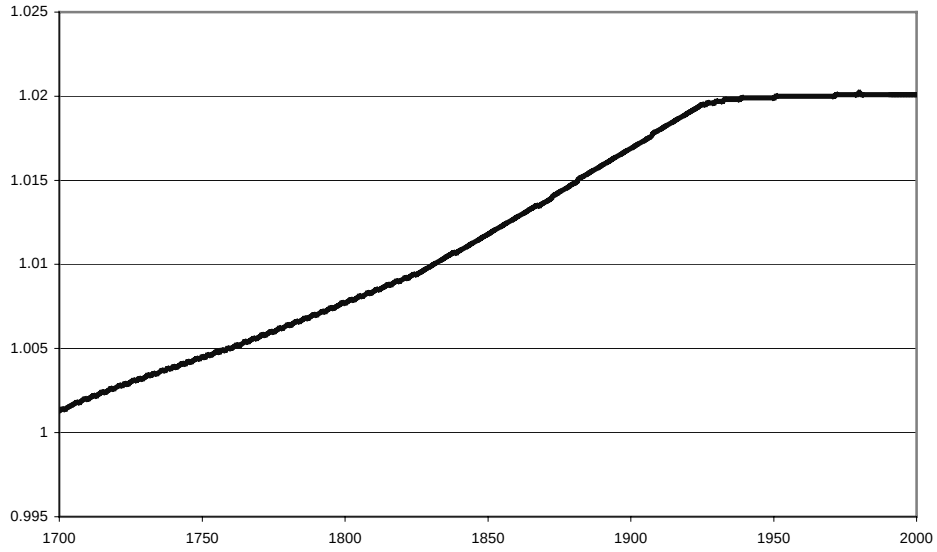


Figure 6: Rental Prices (1700 =1)

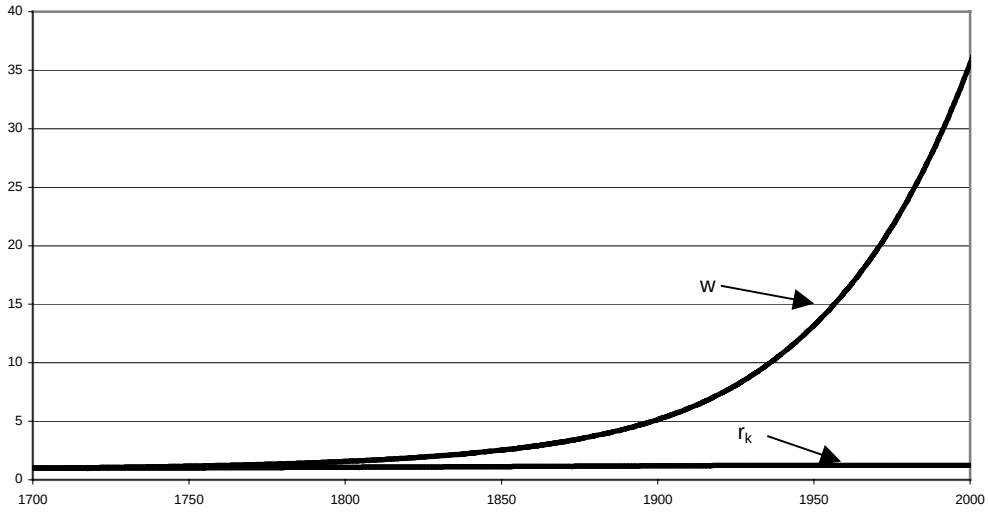


Figure 7: Different Countries Start at Different Times

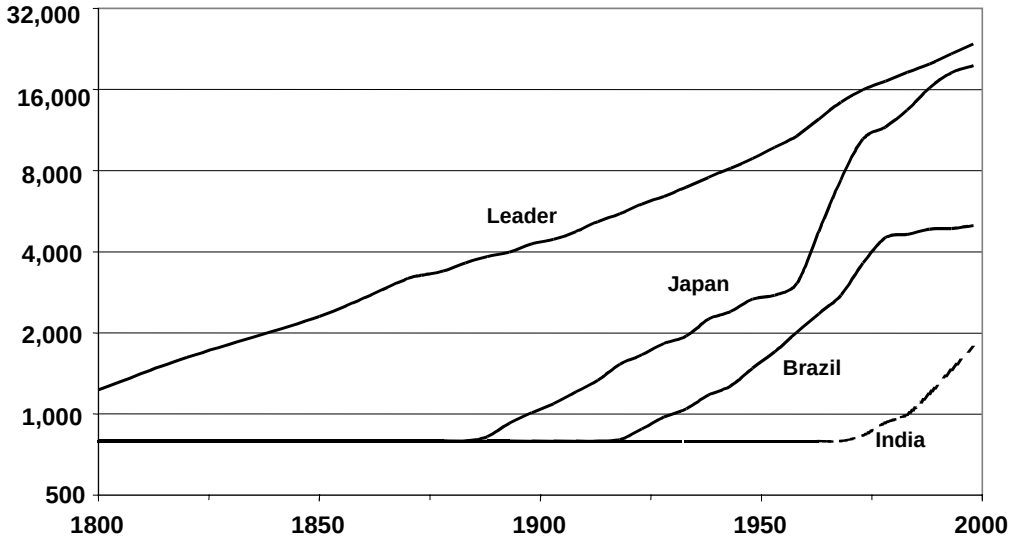


Figure 8: Late Start (Output Relative to the Leader)

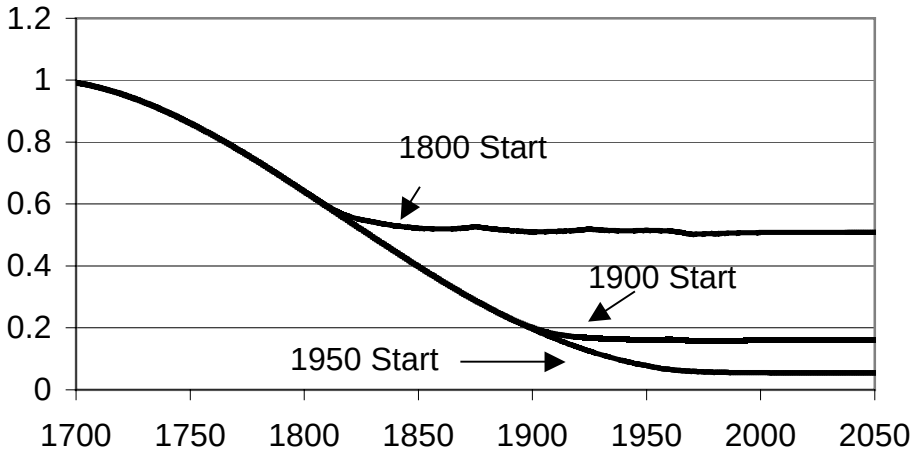


Figure 9: Years for Per Capita Income to Grow from 2,000 to 4,000 (1990 $US)

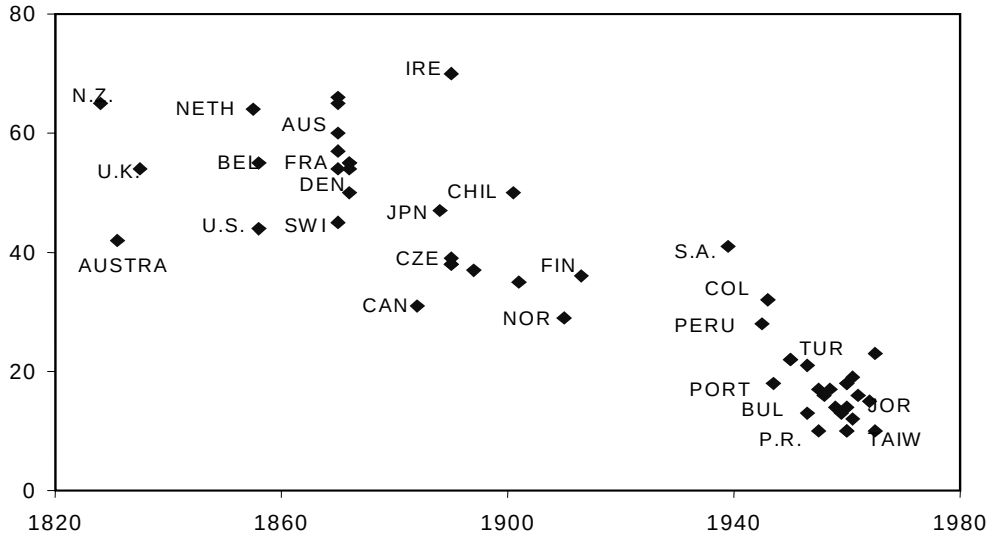


Figure 10: Trends in Output per Capita 1900–95 (1990 $US)

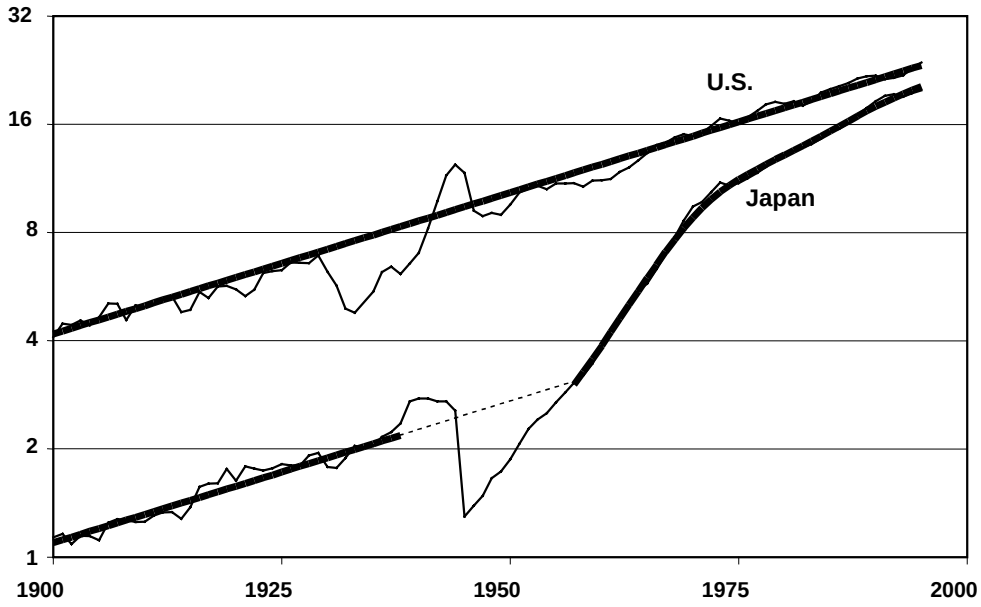


Figure 11: Growth Miracles

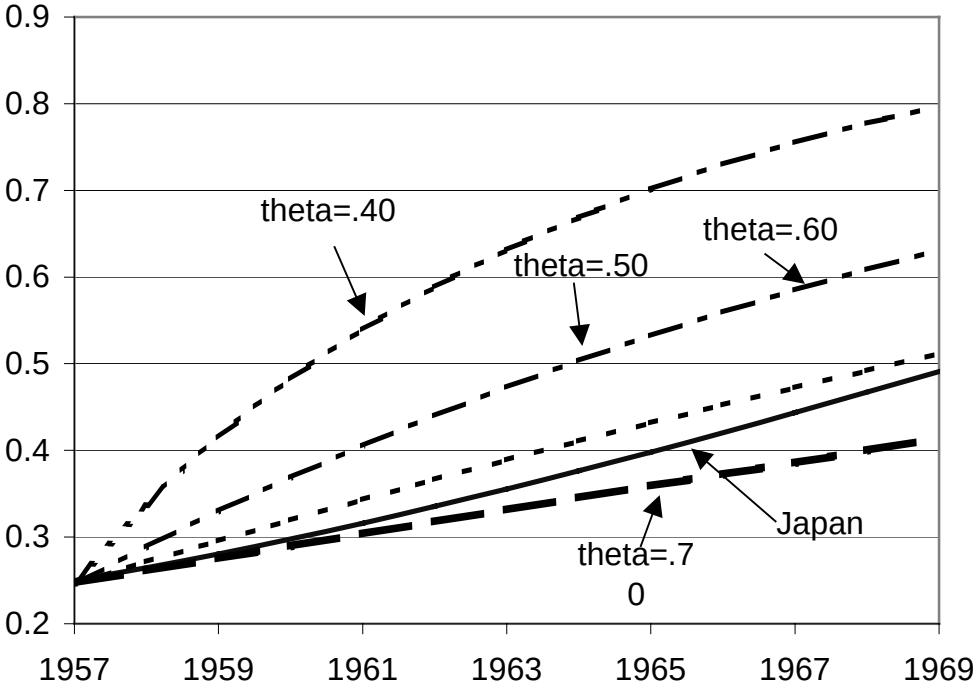
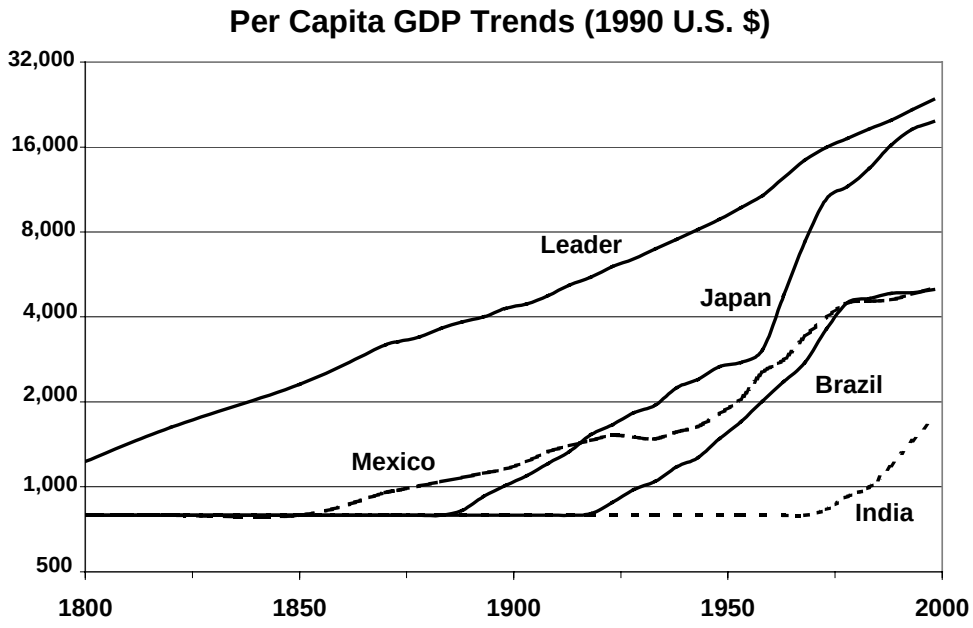


Figure 12

Different Countries Start at Different Times



A Global View Of Economic Growth

Jaume Ventura

CREI and Universitat Pompeu Fabra

March 2005

Abstract: This paper integrates in a unified and tractable framework some of the key insights of the field of international trade and economic growth. It examines a sequence of theoretical models that share a common description of technology and preferences but differ on their assumptions about trade frictions. By comparing the predictions of these models against each other, it is possible to identify a variety of channels through which trade affects the evolution of world income and its geographical distribution. By comparing the predictions of these models against the data, it is also possible to construct coherent explanations of income differences and long-run trends in economic growth.

JEL codes: F10, F15, F40, F43, O11, O40 and O41.

Keywords: Economic growth, international trade, globalization.

I dedicate this research to the memory of Rudi Dornbusch, the best mentor, colleague and friend a young economist could have hoped for, and always capable of making others see things differently. This is a draft of a chapter in the forthcoming *Handbook of Economic Growth* edited by Philippe Aghion and Steven Durlauf. I am thankful to Fernando Broner, Gino Gancia and Francesc Ortega for their useful comments and to Matilde Bombardini, Philip Sauré and Rubén Segura-Cayuela for providing excellent research assistance. I am also grateful to the Fundación Ramón Areces for its generous financial support.

Table of Contents

"All theory depends on assumptions that are not quite true. That is what makes it theory. The art of successful theorizing is to make the inevitable assumptions in such a way that the final results are not very sensitive. A "crucial" assumption is one on which the conclusions do depend sensitively, and it is important that crucial assumptions be reasonably realistic. When the results of a theory seem to flow specifically from a special crucial assumption, then if the assumption is dubious, the results are suspect."

Robert M. Solow [1956, p. 65]

0. Introduction

The world economy has experienced positive growth for an extended period of time. Figure 1 plots average world per capita income from 1500 to today, using data from Maddison's classic study of long run trends in the world economy. The most salient feature of the growth process is its nonlinear nature. For most of the past five hundred years, the world economy settled in a path of stagnation with little growth. But sometime around the early nineteenth century the world economy entered a path of sustained and even accelerating growth. While per capita income grew only by eighteen percent from 1500 to 1820, it has then grown by more than seven hundred and fifty percent from 1820 to today. And this growth has been far from steady. It averaged 0.53 percent from 1820 to 1870, and more than doubled to 1.30 from 1870 to 1913. Growth declined to 0.91 percent during the turbulent period that goes from 1913 to 1950, and then exploded to an unprecedented 2.93 percent from 1950 to 1973. Since then growth has markedly declined to 1.33 percent, even though this period still constitutes the second best growth performance in known human history.

This economic growth has not been distributed equally across the different regions of the world economy. Figure 2 shows per capita income growth for the different regions of the world economy in various time periods. Differences in

regional growth experiences are quite remarkable.[1] Growth took off in Western Europe and its offshoots in the early nineteenth century and never stopped again. But other regions took longer to participate in the growth of the world economy. Perhaps the most dramatic case is that of Asia, which basically did not grow until 1950 just to become then the fastest growing region in the world. Another extreme case is that of Africa, which still today is unable to enjoy growth rates that would be considered modest in other regions. Another salient feature of the growth process is therefore its uneven geographical distribution: in each period there are some regions that have been able to grow and prosper, while others have been left behind.

World economic growth has been accompanied by more than proportional growth in world trade. Figure 3 shows the evolution of world trade as a share of world production since 1870. The picture is quite clear: from 1870 to 1998 growth in world trade has quadrupled growth in world income. There also appears to be a strong positive correlation between growth in per capita income and growth in trade. Figure 4 plots the growth rates of these two variables against each other using pooled data from various regions and periods. The simple correlation between these variables is 0.64, and the regression results indicate that regions and periods with X percent higher than average trade growth tend to have per capita income growth which is 0.3·X higher than average. It almost goes without saying that this statistical association between income and trade does not imply causation in any direction. But it strongly suggests that these variables are somehow related, and that there might be substantial payoffs to working with theories that jointly determine them.[2]

[1] To get a sense of the magnitudes involved, remember that an annual growth rate of G leads per capita income to multiply itself by a factor $F \approx \exp\{G \cdot T\}$ in T years. For instance, in the last quarter of the twentieth century Asia has been able to increase its per capita income by a factor of 2.5, while Latin America has only managed to increase its per capita income by a factor of 1.2 and Africa has stagnated. Even a cursory look at the data shows that this disparity in growth performances constitutes the norm rather than the exception.

[2] For empirical work on the (causal) effect of trade on income levels and income growth see Sachs and Warner [1995], Frankel and Romer [1999], Ades and Glaeser [1999], Alesina, Spolaore, and Wacziarg [2000 and their chapter in this handbook], Rodriguez and Rodrik [2000], Alcalá and Ciccone [2003 and 2004], and Dollar and Kraay [2003].

Despite this apparent relationship between income and trade, a substantial part of growth theory is built on the assumption that countries live in autarky and that there is no trade among them.[3] This is obviously a dubious assumption. But is it also a "crucial" one? And if so, what alternative assumptions would be reasonably realistic? At an abstract level, these are the questions that I attempt to answer here. A recurring theme throughout this chapter is that the growth experiences of the different world regions are intimately linked and cannot be analyzed in isolation. We therefore need a global view of economic growth that looks at the different regions of the world as parts of a single whole. Formally, this means that we should develop and systematically study world equilibrium models. These models and their predictions constitute the specific focus of this chapter. [4]

Rather than providing an all-encompassing survey of the field, my goal in writing this chapter has been to develop a unified and yet tractable framework to discuss key insights of the fields of international trade and economic growth. In particular, I examine a sequence of world equilibrium models that share a common description of technology and preferences but differ on their assumptions about trade frictions. By comparing the predictions of these models against each other, it is possible to identify a variety of channels through which trade affects the evolution of world income and its geographical distribution. By comparing their predictions against the data, it is also possible to construct coherent explanations of income differences and long run trends in economic growth. When viewed as a group, these models show that much is known about the relationship between income and trade. Despite this, I still feel we are only exploring the tip of the iceberg. The research program sketched here is ambitious, fun and it could eventually lead to a much deeper understanding of the forces that drive modern capitalist economies.

[3] A brief examination of the different chapters of this handbook should quickly convince anyone doubting this statement.

[4] Without doubt, the seminal book by Grossman and Helpman [1991] is the single most influential contribution to the development and study of world equilibrium models of the growth process. It heavily influenced a whole generation of PhD students, like myself, that were searching for dissertation topics when the book first appeared. But there are, of course, many other important contributions. The bibliography at the end of this chapter is an (admittedly imperfect) attempt to list all published papers that use world equilibrium models to study the growth process. I apologize to the authors of any relevant paper that has been overlooked.

The rest of this chapter contains four sections. The first one describes growth in the integrated economy. This is an imaginary world where trade costs are negligible and geography does not matter. Section two introduces two trade frictions: the immobility of production factors and the absence of international financial markets. Section three adds a third trade friction: costs of transporting goods. The fourth and final section briefly concludes by taking stock what we have learned and pointing out potential avenues for further research.

1. The integrated economy

Imagine a world without borders, a world in which all goods and factors can be transported across different regions at negligible cost. Some industries spread their production process across many regions searching for the ideal environment for each specific phase of production. Other industries choose instead to concentrate production in a single region to exploit increasing returns to scale. Regardless of an industry's particular circumstances, its location choice maximizes productivity and is not affected by the local availability of production factors and/or final customers. If a region does not have the necessary production factors, these can be imported from abroad. If a region does not have enough customers, the goods produced can be exported abroad. In this world, global market forces arbitrage away regional differences in goods and factor prices and all the gains from trade are reaped. This imaginary world is the integrated economy, and is the subject of this section.

The integrated economy provides a natural benchmark for the study of economic growth in an interdependent world. Moreover, its simplicity and elegance encapsulates the essence of what growth theory is all about: deriving strong results using minimalist models. In the spirit of the so-called "new growth theory", I shall use a model that jointly determines the stock of capital and the level of technology.

Admittedly, the model is somewhat lopsided. On the one hand, it contains a fairly sophisticated formulation of technology that includes various popular models as special cases. On the other hand, it uses a brutal simplification of the standard overlapping-generations model as a description of preferences. Despite this, I do not apologize for the imbalance. A robust theme in growth theory is that the interesting part of the story is nearly always on the technology side, and rarely on the side of preferences.

This section develops the basic framework that I use throughout the chapter. Sub-section 1.1 describes the integrated economy, while sub-section 1.2 derives its main predictions for world growth. Sub-section 1.3 goes back to a period in which all the regions of the world lived in autarky, and compares the growth process of this world with the integrated economy. This is just the first of various attacks to the question of globalization and its effects on the world economy.

1.1 A workhorse model

Consider a world economy inhabited by two overlapping generations: young and old. The young work and, if productive, they earn a wage. The old retire and live off their savings. All generations have size one. There are many final goods used for consumption and investment, indexed by $i \in I$. When this does not lead to confusion, I shall use I to refer both to the set of final goods and also to the number of final goods. As we shall see later, the production of these final goods requires a continuum of intermediate inputs. There are two factors of production: labor and capital. For simplicity, I assume capital depreciates fully within one generation.[5] The world economy contains many regions. But geography has no economic consequences since goods and factors can be transported from one region to another at any time at negligible cost.

[5] The main role of this assumption is to ensure that investment is always strictly positive. This simplifies the presentation without substantially affecting the main results.

The citizens of this world differ in their preferences and access to education. S_t members of the generation born in date t are patient and maximize the expected utility of old age consumption, while the rest are impatient and maximize the expected utility of consumption when young. The utility function has consumption as its single argument, and it is homothetic, strictly concave and identical for all individuals. H_t members of the generation born in date t can access education and become productive, while the rest have no access to education and remain unproductive.[6] I refer to S_t and H_t as "savings" and "human capital", and I allow them to vary stochastically over time within the unit interval. Assuming that savings and human capital are uncorrelated within each generation, we obtain:

$$(1) \qquad K_{t+1} = S_t \cdot w_t \cdot H_t$$

$$(2) \qquad C_t = (1 - S_t) \cdot w_t \cdot H_t + r_t \cdot K_t$$

where K_t and C_t are the average or aggregate capital stock and consumption; and w_t and r_t are the wage and rental rate of capital. Equation (1) states that the capital stock equals the savings of the young, which consist of the wage of those that are patient and productive. The assumption that capital depreciates fully in one generation implies that the capital stock is equal to investment. Equation (2) says that consumption equals the wage of the impatient and productive young plus the return to the savings of the old.[7]

Consumption and investment can be thought of as composites or aggregates of the different final goods. A very convenient assumption is that both composites

[6] The assumption that labor productivity is either one or zero is extreme, but inessential. We could also think of H_t as the average labor productivity of the world economy. The assumption that human capital is not industry specific is widespread, but not entirely innocent. See Basu and Weil [1998] and Brezis, Krugman and Tsiddon [1993] for interesting implications of relaxing this assumption.

[7] This representation of savings and consumption is nothing but a stripped-down version of Modigliani's life-cycle theory of savings. It abstracts from other motives for savings such as leaving bequests. These could be easily re-introduced in the theory through suitable and well-known modifications of the preferences of individuals. I shall not do this to keep the analysis as simple as possible. I conjecture that the bulk of the basic intuitions and results presented here would not be meaningfully affected by these extensions.

take the same Cobb-Douglas form with spending shares that vary across industries, i.e. σ_i with $\sum_{i \in I} \sigma_i = 1$. Since there is a common ideal price index for consumption and investment, it makes sense to use it as the numeraire and this implies that aggregate spending is given by $E_t \equiv C_t + K_{t+1}$. To sum up, we have that:

(3) $E_{it} = \sigma_i \cdot E_t$ for all $i \in I$

(4) $1 = \prod_{i \in I} \left(\frac{P_{it}}{\sigma_i} \right)^{\sigma_i}$

where E_{it} and P_{it} are the total spending on and the price of the final good of industry i. Equation (3) states that spending shares are constant, while Equation (4) sets the common price of consumption and investment equal to one.

Production of final goods uses labor, capital and a continuum of different varieties of intermediate inputs, indexed by $m \in [0, M_{it}]$ for all $i \in I$. As usual, I interpret the measure of input varieties, M_{it} for all $i \in I$, as the degree of specialization or the technology of the industry. This measure will be determined endogenously as part of the equilibrium. The technology of industry i can be summarized by these total cost functions:

(5) $B_{it} = \left[\frac{1}{Z_{it}} \cdot \left(\frac{w_t}{1-\alpha_i} \right)^{1-\alpha_i} \cdot \left(\frac{r_t}{\alpha_i} \right)^{\alpha_i} \right]^{1-\beta_i} \cdot \left[\int_0^{M_{it}} p_{it}(m)^{1-\varepsilon_i} \cdot dm \right]^{\frac{\beta_i}{1-\varepsilon_i}} \cdot Q_{it}$ for all $i \in I$

(6) $b_{it}(m) = \frac{1 + q_{it}(m)}{Z_{it}} \cdot \left(\frac{w_t}{1-\alpha_i} \right)^{1-\alpha_i} \cdot \left(\frac{r_t}{\alpha_i} \right)^{\alpha_i}$ for all $m \in [0, M_{it}]$ and $i \in I$

where $0 \leq \beta_i \leq 1$, $\varepsilon_i > 1$ and $0 \leq \alpha_i \leq 1$; Q_{it} is total production of final good i; $q_{it}(m)$ and $p_{it}(m)$ are the quantity and price of the m^{th} input variety of industry i; and the variables Z_{it}

are meant to capture the influence on industry productivity of geography, institutions and other factors that are exogenous to the analysis.[8] I loosely refer to the Z_{it}s as "industry productivities" and assume they vary stochastically over time within a support that is strictly positive and bounded above. Equation (5) states that the technology to produce the final good of industry i is a Cobb-Douglas function on human and physical capital, and intermediate inputs. The latter are aggregated with a standard CES function. Equation (6) states that the production of intermediates is also a Cobb-Douglas function on human and physical capital, and that there are fixed and variable costs.[9] I interpret the fixed costs as including both the costs of building a specialized production plant and the costs of inventing or developing a new variety of intermediate. An important simplifying assumption is that input varieties become obsolete in one generation and, as a result, all generations must incur these fixed costs.[10]

Since there are constant returns in the production of final goods, it is natural to assume that final good producers operate under perfect competition. Therefore, prices and intermediate input demands are given as follows:

$$(7) \qquad P_{it} = \frac{\partial B_{it}}{\partial Q_{it}} \qquad \text{for all } i \in I$$

$$(8) \qquad q_{it}(m) = \frac{\partial B_{it}}{\partial p_{it}(m)} \qquad \text{for all } m \in [0, M_{it}] \text{ and } i \in I$$

Equation (7) states that price equals marginal cost, while Equation (8) uses Shephard's lemma to describe the demand for intermediate inputs. Equations (5) and (8) imply that an increase in the price of a given input variety lowers its market share. But Equation (3) shows that the lost market share goes entirely to other input

[8] Although popular, this is a quite simplistic view of the effects of geography and institutions. See Levchenko [2004] for an interesting discussion of alternative ways of modeling the effects of institutions.
[9] As usual, the fixed cost is paid if and only if there is strictly positive production.
[10] This assumption is crucial for tractability, since it eliminates a potentially large set of state variables, i.e. M_{it} for all $i \in I$.

varieties of the same industry and does not affect the industry's overall market share.

Since the production of intermediate inputs exhibits increasing returns that are internal to the firm, input producers cannot operate under perfect competition. I assume instead they operate under monopolistic competition with free entry. This has the following implications:

(9) $$p_{it}(m) = \frac{e_{it}(m)}{e_{it}(m) - 1} \cdot \frac{\partial b_{it}(m)}{\partial q_{it}(m)} \qquad \text{for all } m \in [0, M_{it}] \text{ and } i \in I$$

(10) $$p_{it}(m) \cdot q_{it}(m) = b_{it}(m) \qquad \text{for all } m \in [0, M_{it}] \text{ and } i \in I$$

where $e_{it}(m)$ is the price-elasticity of input demand: $e_{it}(m) = -\frac{p_{it}(m)}{q_{it}(m)} \cdot \frac{\partial q_{it}(m)}{\partial p_{it}(m)}$ with the derivative in this definition being applied to Equation (8). Equation (9) states that monopolistic firms charge a markup over marginal cost that is decreasing on the demand elasticity faced by the firm. As usual, the CES formulation implies that this demand elasticity is equal to the elasticity of substitution among inputs, i.e. $e_{it}(m)=\varepsilon_i$. Equation (10) states that profits must be zero and this is, of course, a direct implication of assuming free entry.

Finally, we must impose appropriate resource constraints or market-clearing conditions:

(11) $$P_{it} \cdot Q_{it} = E_{it} \qquad \text{for all } i \in I$$

(12) $$H_t = \sum_{i \in I} H_{it} \qquad \text{with} \quad H_{it} = \frac{\partial B_{it}}{\partial w_t} + \int_0^{M_{it}} \frac{\partial b_{it}(m)}{\partial w_t} \cdot dm$$

(13) $$K_t = \sum_{i \in I} K_{it} \qquad \text{with} \quad K_{it} = \frac{\partial B_{it}}{\partial r_t} + \int_0^{M_{it}} \frac{\partial b_{it}(m)}{\partial r_t} \cdot dm$$

where H_{it} and K_{it} are the labor and capital demanded by industry i. Since the integrated economy is a closed economy, Equation (11) forces the aggregate supply of each good to match its demand, while Equations (12)-(13) state that the aggregate supply of labor and capital must equal their demands. The latter are the sum of their industry demands, and these are calculated using Shephard's lemma.

This completes the description of the model. For any admissible initial capital stock and sequences for S_t, H_t, and Z_{it}, an equilibrium of the integrated economy consists of sequences of prices and quantities such that Equations (1)-(13) hold in all dates and states of nature. The assumptions made ensure that this equilibrium always exists and is unique. I shall show this by construction in the next section.

The reader might be wondering why I have not formally introduced financial markets. I have allowed individuals to construct their own capital and use it as a vehicle to carry on their savings into retirement (a world of family-owned firms?). But I have not allowed them to trade securities in organized financial markets. The reason is simply to save notation. The assumptions made ensure that asset trade does not matter in this world economy. [11] To see this, assume there exist sophisticated financial markets where all individuals can trade a wide array of state-contingent securities. Naturally, the old would not be able to trade these securities since they will not be back to settle claims one period later. But the young would not trade with each other either. Impatient young would not be willing to trade securities since they do not have income in their old age and are happy to consume all their income during their youth. Patient young are the only ones willing and able to trade these securities. But they all have identical preferences and face the same distribution of returns to capital, and therefore they find no motive to trade with each other. Thus, we can safely assume the integrated economy contains sophisticated

[11] This statement is not entirely correct. It applies to assets whose price reflects only fundamentals, but without additional assumptions it does not apply to securities whose price contains a bubble. I shall disregard the possibility of asset bubbles in this chapter, although this is far from an innocuous assumption. See Ventura [2002] for an example where asset bubbles have an important effect on the growth of the world economy and its geographical distribution.

financial markets that allow individuals to enter contracts that specify exchanges of various quantities of the different goods to be delivered at various dates and/or states of nature. It just happens that these financial markets do not make any difference for consumption and welfare.

1.2 Diminishing returns, market size and economic growth

To study the forces that determine economic growth in the integrated economy, it is useful to start with a familiar expression:

$$(14) \qquad \frac{K_{t+1}}{K_t} = s_t \cdot \frac{Q_t}{K_t}$$

where Q_t is the integrated economy's output or production, i.e. $Q_t \equiv \sum_{i \in I} P_{it} \cdot Q_{it}$; and s_t is the economy's (gross) savings rate, i.e. $s_t \equiv \frac{K_{t+1}}{Q_t}$. Equation (14) states that the (gross) growth rate of the capital stock is equal to the savings rate times the output-capital ratio or average product of capital. If this product stays above one asymptotically, the world economy exhibits sustained or long run growth. Otherwise, economic growth eventually ceases and the world economy stagnates. We shall study then the determinants of savings and the average product of capital.

To compute the savings rate, remember that industry i receives a share σ_i of aggregate spending of which a fraction $1-\alpha_i$ goes to labor. Adding across industries, it follows that aggregate labor income is $w_t \cdot H_t = (1-\alpha) \cdot Q_t$, where α is the aggregate or average share of capital, i.e. $\alpha \equiv \sum_{i \in I} \sigma_i \cdot \alpha_i$. Since only the patient young save, the savings rate consists of the fraction of labor income in the hands of patient consumers:

(15) $s_t = (1-\alpha)\cdot S_t$

Since the savings rate is bounded above, sustained economic growth requires that the average product of capital remain above one as the economy grows. But what determines the aggregate output-capital ratio? I shall answer this question in a few steps, so as to develop intuition.

The first step consists of finding the output-capital ratio of a given industry as a function of its technology and factor proportions:[12]

(16) $$\frac{Q_{it}}{K_{it}} = \left(\frac{\varepsilon_i}{\varepsilon_i - 1}\right)^{-\beta_i} \cdot M_{it}^{\frac{\beta_i}{\varepsilon_i - 1}} \cdot Z_{it} \cdot \left(\frac{K_{it}}{H_{it}}\right)^{\alpha_i - 1} \qquad \text{for all } i \in I$$

Equation (16) shows the effects of changes in factor proportions on the industry's output-capital ratio, *holding constant technology*. Since there are diminishing returns to physical and human capital in production, we find the standard result that increases in the physical to human capital ratio reduce the output-capital ratio. But technology is endogenously determined in this model, and it depends on the size of the industry:[13]

(17) $$M_{it} = \frac{\beta_i}{\varepsilon_i} \cdot Z_{it} \cdot H_{it}^{1-\alpha_i} \cdot K_{it}^{\alpha_i} \qquad \text{for all } i \in I$$

Equation (17) shows that increases in factor usage or industry size raise the incentives to specialize and therefore improve technology. The larger is the size of the market, the easier it is to recoup the fixed costs of producing a new input variety

[12] From Equations (7) and (11) find that $P_{it}\cdot Q_{it}=B_{it}$, and use this to eliminate B_{it} from Equation (5). Then, solve Equation (9) with Equation (6), substitute into Equation (5) and eliminate factor prices by noting that the industry factor shares, i.e. $w_t\cdot H_{it}/P_{it}\cdot Q_{it}$ and $r_t\cdot K_{it}/P_{it}\cdot Q_{it}$ are given by $1-\alpha_i$ and α_i, respectively.

[13] Symmetry of intermediates and perfect competition in the final goods industry implies that $M_{it}\cdot p_{it}\cdot q_{it}=\beta_i\cdot P_{it}\cdot Q_{it}$; where p_{it} and q_{it} are the common price and quantity of all varieties of intermediates of industry i. Then, use Equations (6), (9) and (10) to eliminate p_{it} and q_{it} from this expression. Finally, eliminate factor prices once again by noting that the industry factor shares are $1-\alpha_i$ and α_i.

and therefore the higher is the number of input varieties that can be sustained in equilibrium. We can now put these two pieces together and write the output-capital ratio as follows:

$$(18) \qquad \frac{Q_{it}}{K_{it}} = A_{it} \cdot H_{it}^{\mu_i \cdot (1-\alpha_i)} \cdot K_{it}^{\mu_i \cdot \alpha_i - 1} \qquad \text{for all } i \in I$$

where μ_i is a measure of the importance of market size effects, i.e. $\mu_i = 1 + \frac{\beta_i}{\varepsilon_i - 1}$;

and A_{it} is a measure of industry productivity, i.e. $A_{it} = \left(\frac{\varepsilon_i}{\varepsilon_i - 1}\right)^{-\beta_i} \cdot \left(\frac{\beta_i}{\varepsilon_i}\right)^{\frac{\beta_i}{\varepsilon_i - 1}} \cdot Z_{it}^{\mu_i}$. I shall refer to both Z_{it} and A_{it} as "industry productivities" when this is not a cause for confusion. Equation (18) summarizes the aggregate industry technology and shows direct and indirect effects of factor usage on the industry's output-capital ratio. Increases in human capital raise the output-capital ratio, as the direct positive effect of making physical capital scarce is reinforced by the indirect effect of increasing input variety. Increases in physical capital have an ambiguous effect on the output-capital ratio, as the direct negative effect of making physical capital abundant and the positive indirect effect of increasing input variety work in opposite directions. If diminishing returns are strong and market size effects are weak ($\mu_i \cdot \alpha_i < 1$) increases in physical capital reduce the industry's output-capital ratio. If instead diminishing returns are weak and market size effects are strong ($\mu_i \cdot \alpha_i \geq 1$) increases in physical capital raise the industry's output-capital ratio.

The next step is to aggregate these effects across industries. To do this, note first that factor allocations and aggregate output are determined as follows:[14]

[14] Equations (19) and (20) are direct implications of the constant factor and spending shares. One way to think about Equation (21) is as the definition of the Cobb-Douglas aggregate that defines consumption and investment and therefore underlies Equations (3) and (4). Another way of thinking about Equation (21) is as an implication of Equations (3), (4) and (11).

(19) $H_{it} = \sigma_i \cdot \frac{1-\alpha_i}{1-\alpha} \cdot H_t$ for all $i \in I$

(20) $K_{it} = \sigma_i \cdot \frac{\alpha_i}{\alpha} \cdot K_t$ for all $i \in I$

(21) $Q_t = \prod_{i \in I} Q_{it}^{\sigma_i}$

Equations (19) and (20) show that the equilibrium allocations of human and physical capital to industry i depend on the corresponding factor share and the size of the industry. Equation (21) says that output is a Cobb-Douglas aggregate of industry outputs. This is, of course, the production function associated with the cost function in Equation (4). It is now immediate to substitute Equations (18), (19) and (20) into Equation (21) to find the aggregate output-capital ratio of the world economy:

(22) $\frac{Q_t}{K_t} = A_t \cdot H_t^{\mu\cdot(1-\alpha)-\upsilon} \cdot K_t^{\mu\cdot\alpha+\upsilon-1}$

where μ is the average value of μ_i, i.e. $\mu \equiv \sum_{i \in I} \sigma_i \cdot \mu_i$; υ is the covariance between μ_i and α_i, i.e. $\upsilon \equiv \sum_{i \in I} \sigma_i \cdot (\mu_i - \mu) \cdot (\alpha_i - \alpha)$; and A_t is an aggregate measure of productivity, i.e. $A_t \equiv \prod_{i \in I} \left[\sigma_i^{\mu_i} \cdot \left(\frac{1-\alpha_i}{1-\alpha} \right)^{\mu_i \cdot (1-\alpha_i)} \cdot \left(\frac{\alpha_i}{\alpha} \right)^{\mu_i \cdot \alpha_i} \cdot A_{it} \right]^{\sigma_i}$. Equation (22) is the aggregate production function and will play an important role in what follows. It shows that the industry intuitions on the effects of changes in factor usage carry on to the aggregate effects of changes in factor supplies. While increases in human capital unambiguously raise the output-capital ratio, increases in physical capital have ambiguous effects.[15] If the "representative" industry has strong diminishing returns and weak market-size effects ($\mu\cdot\alpha+\upsilon<1$) physical capital accumulation

[15] Note that $\mu\cdot(1-\alpha)-\upsilon \geq 0$.

reduces the aggregate output-capital ratio. If instead the "representative" industry has weak diminishing returns and strong market-size effects ($\mu\cdot\alpha+\upsilon\geq 1$) physical capital accumulation raises the output-capital ratio.

We are ready now to characterize the process of economic growth in the integrated economy. Substituting Equation (22) into Equation (14), we obtain the following law of motion for the capital stock:

$$(23) \qquad K_{t+1} = s_t \cdot A_t \cdot H_t^{\mu\cdot(1-\alpha)-\upsilon} \cdot K_t^{\mu\cdot\alpha+\upsilon}$$

Equation (23) shows that the integrated economy behaves as if it were a Solow model with a Cobb-Douglas production function that exhibits increasing returns to scale, i.e. the sum of the share coefficients is $\mu\geq 1$. Figures 5 and 6 illustrate the dynamics of the stock of physical capital with the help of two simple examples. The first example is the "deterministic" world where savings, human capital and productivity are constant over time, i.e. $\{s_t,H_t,A_t\}=\{s,H,A\}$ for all t. The second example is the "stochastic" world where savings, human capital and productivity fluctuate between a "bad" state with $\{s_t,H_t,A_t\}=\{s_B,H_B,A_B\}$ and a "good" state with $\{s_t,H_t,A_t\}=\{s_G,H_G,A_G\}$; with $s_G \cdot A_G \cdot H_G^{\mu\cdot(1-\alpha)-\upsilon} > s_B \cdot A_B \cdot H_B^{\mu\cdot(1-\alpha)-\upsilon}$. The central point of these examples is to show that economic growth solves a tension between diminishing returns and market size effects.

Figure 5 shows the case in which diminishing returns are strong and market-size effects are weak, i.e. $\mu\cdot\alpha+\upsilon<1$. The top panel depicts the evolution of the "deterministic" world. There is a unique steady state and the stock of physical capital converges monotonically towards it from any initial position. The steady state is stable because increases (decreases) in the stock of physical capital lower (raise) the output-capital ratio and lead to a lower (higher) growth rate. The bottom panel shows that the "stochastic" world exhibits similar dynamics, with the stock of physical capital monotonically converging to a steady state interval, rather than a steady state

value. Once the stock of physical capital is trapped within this interval, its growth rate fluctuates between positive and negative values and averages zero in the long run. These examples illustrate why sustained growth is not possible if diminishing returns are strong and market size effects are weak.

Figure 6 shows the case in which diminishing returns are weak and market-size effects are strong, i.e. $\mu \cdot \alpha + \upsilon \geq 1$. The top panel shows the "deterministic" world again. There is unique steady state that is unstable. If the stock of physical capital starts above the steady state, it grows without bound at an accelerating rate. If it starts below, the stock of physical capital contracts over time also at an accelerating rate. The steady state is now unstable because increases (decreases) in the stock of physical capital raise (lower) the output-capital ratio and lead to a higher (lower) growth rate. The bottom panel shows that the "stochastic" world also exhibits similar dynamics. One difference however is that there is no steady state. Instead, there is a threshold interval. If the stock of physical capital is above (below) this interval, it grows (contracts) at an accelerating rate. If the stock of physical capital starts within the threshold interval, it fluctuates within it until it eventually exits. This happens with probability one, and only luck determines when this exit occurs and whether the world economy exits above and enters an expansionary path or, alternatively, it exits below and enters a contractionary path. Therefore, sustained growth is possible (but not necessary) if diminishing returns are weak and market size effects are strong.

This model suggests a simple account of the history of the world economy since the 1500s. It is based on the "stochastic" world of Figure 6 and it goes as follows: for centuries, the size of the world economy was too small to generate sustained growth. Located within the threshold interval, the world economy was subject to periodic expansions and contractions with virtually zero average growth. This is consistent with Maddison's calculation that the world economy grew only about eighteen percent from 1500 to 1820. But this was an unstable situation in the very long run. The Industrial Revolution marks the moment in which, after a series of favorable shocks, the world economy reached enough size to exit the threshold

interval and started traveling on the path of accelerating growth reported in Figure 1. As a result of this successful exit, the world economy grew more than seven hundred and fifty percent from 1820 to 1998.

Although suggestive, this account is far too sketchy and incomplete to be taken seriously. Moreover, I find highly improbable that the last five hundred years of the world economy can be understood in terms of a model that postulates negligible costs of transporting goods and factors and constant world population. Surely the demographic revolution and the process of globalization have both played central roles in shaping the growth process during this period. This chapter is not the place for a discussion of the growth effects of the demographic revolution.[16] But it is definitely the place to study the growth effects of globalization, and we turn to this topic next.

1.3 The effects of economic integration

Assume the world economy initially consisted of many regions or locations separated by geographical obstacles that made the costs of transporting goods and factors among them prohibitive. As a result, these regions were forced to live in autarky. I index these regions by $c \in C$, and let them differ on their savings, human capital, industry productivities and initial capital stock, i.e. on $S_{c,t}$, $H_{c,t}$, $Z_{c,it}$ and $K_{c,0.}$ When this does not lead to confusion, I shall use C to refer to both the set of regions and also to the number of regions. Throughout, I denote world aggregates by omitting the region sub-index. Typically, world aggregates refer to the sum of all corresponding regional variables. For instance, world aggregate savings, human and

[16] In this model, a sustained increase in population would generate sustained growth even if $\alpha \cdot \mu + \upsilon < 1$. The reason is that, holding constant both factor endowments and productivity, population growth increases the size of the market and this raises income. I have ruled out this possibility by simply assuming that the world population is constant. Given the purpose of this chapter, I think this is not a "crucial" assumption. But it might be so in other contexts. See Jones' chapter in this volume for a thorough and clear discussion of scale effects in growth models.

physical capital are $S_t = \sum_{c \in C} S_{c,t}$, $H_t = \sum_{c \in C} H_{c,t}$ and $K_t = \sum_{c \in C} K_{c,t}$. But there will be some exceptions. For instance, the relationship between $Z_{c,it}$ and the corresponding world aggregate Z_{it} is a bit more intricate and will be explained shortly.

Although it is not really necessary to take a stand on the geographical distribution of population, I assume throughout that it is equally distributed across regions. This simplifies somewhat the presentation since absolute and per capita regional comparisons coincide. For instance, if $S_{c,t}>S_{c',t}$ then c also has higher savings per person than c'. Note also that, as the number of regions becomes arbitrarily large, the size of each of them becomes arbitrarily small and the effects of shocks to their characteristics on world aggregates become arbitrarily small. This limiting case is usually referred to as the small economy assumption.

The model of globalization considered here is embarrassingly simple: at date t=0, all the geographical obstacles to trade suddenly disappear forever and the costs of transporting goods and factors fall from prohibitive to negligible. What are the effects of such a dramatic reduction in transport costs on world economic growth and its geographical distribution? To answer this question, we must characterize the growth process in the autarkic world economy and in the integrated world economy and compare them. Although this way of modeling globalization and its effects is almost a caricature, it turns out to be quite useful to develop intuitions that survive as we move to more sophisticated and realistic models.

In the world of autarky, each region constituted a smaller version of the integrated economy. Therefore, the world economy at t<0 can be described by:[17]

(24) $Y_{c,t} = A_{c,t} \cdot H_{c,t}^{\mu\cdot(1-\alpha)-\upsilon} \cdot K_{c,t}^{\mu\cdot\alpha+\upsilon}$ for all $c \in C$

(25) $K_{c,t+1} = s_{c,t} \cdot A_{c,t} \cdot H_{c,t}^{\mu\cdot(1-\alpha)-\upsilon} \cdot K_{c,t}^{\mu\cdot\alpha+\upsilon}$ for all $c \in C$

[17] Equation (25) is an analogue to Equation (23), while Equation (24) follows from the region counterparts to Equation (22) and the fact that $Y_{c,t}=Q_{c,t}=C_{c,t}+K_{c,t}$ in autarky.

where $Y_{c,t}$ is the income of the region and, in autarky, it coincides with its production and spending, i.e. $Y_{c,t}=Q_{c,t}=E_{c,t}$; and $A_{c,t}$ is the corresponding measure of regional productivity, i.e. $A_{c,t} \equiv \prod_{i \in I}\left[\sigma_i^{\mu_i} \cdot \left(\frac{1-\alpha_i}{1-\alpha}\right)^{\mu_i \cdot (1-\alpha_i)} \cdot \left(\frac{\alpha_i}{\alpha}\right)^{\mu_i \cdot \alpha_i} \cdot A_{c,it}\right]^{\sigma_i}$ with $A_{c,it} = \left(\frac{\varepsilon_i}{\varepsilon_i - 1}\right)^{-\beta_i} \cdot \left(\frac{\beta_i}{\varepsilon_i}\right)^{\frac{\beta_i}{\varepsilon_i - 1}} \cdot Z_{c,it}^{\mu_i}$. Equations (24) and (25) have been discussed at length already and need no further comment.

In the integrated economy it is not possible in general to determine the production or spending located in a given region. Since goods and factors can move at negligible cost, any geographical distribution of production and factors that ensures all production takes place in the regions with the highest industry productivity is a possible equilibrium. Despite this indeterminacy, prices and aggregate quantities are uniquely determined as shown in section 1.2. This means that it is possible to track the stock of physical capital owned by the original inhabitants of region c and their descendants as well as their income:[18]

(26) $$Y_{c,t} = \left[(1-\alpha) \cdot \frac{H_{c,t}}{H_t} + \alpha \cdot \frac{K_{c,t}}{K_t}\right] \cdot A_t \cdot H_t^{\mu \cdot (1-\alpha) - \upsilon} \cdot K_t^{\alpha \cdot \mu + \upsilon} \qquad \text{for all } c \in C$$

(27) $$K_{c,t+1} = \frac{S_{c,t} \cdot H_{c,t}}{S_t \cdot H_t} \cdot s_t \cdot A_t \cdot H_t^{\mu \cdot (1-\alpha) - \upsilon} \cdot K_t^{\alpha \cdot \mu + \upsilon} \qquad \text{for all } c \in C$$

for all $c \in C$ and $t \geq 0$; and A_t is a measure of world productivity. Remember that we have now specified a set of industry productivities for each region, $Z_{c,it}$. But we only specified one set of industry productivities for the integrated economy in section 1.1.

[18] Equation (26) follows from adding the income from human and physical capital of the inhabitants of the region, and noting that aggregate or world shares of human and physical capital are constant and equal to 1-α and α, respectively. Equation (27) follows from Equations (1) and (23), and the observation that wages are the same for all productive workers of the world. Without loss of generality, I keep assuming that there is no trade in securities.

The reason was that industries never locate in a region that offers less than the highest possible productivity. As a result, in the integrated world economy the only industry productivities that matter are the highest ones, i.e. $Z_{it} = \max_{c \in C}\{Z_{c,it}\}$. This implies that $A_t \geq A_{c,t}$ for all $c \in C$, and we can interpret aggregate productivity not as average productivity, but instead as the highest possible productivity or the world productivity frontier. With this in mind, Equation (27) traces the holdings of capital of the original inhabitants of region c and their descendants, while Equation (26) describes their income.

We are ready now to examine the growth effects of economic integration. Consider first the static or impact effects on the incomes of regions. A bit of straightforward algebra shows that:[19]

$$(28) \quad \ln\left(\frac{Y^{I}_{c,0}}{Y^{A}_{c,0}}\right) = \underbrace{\ln\left(\frac{A_0}{A_{c,0}}\right)}_{\substack{higher \\ productivity}} + \underbrace{\ln\left(\frac{(1-\alpha)\cdot\frac{H_{c,0}}{H_0} + \alpha\cdot\frac{K_{c,0}}{K_0}}{\left(\frac{H_{c,0}}{H_0}\right)^{1-\alpha}\cdot\left(\frac{K_{c,0}}{K_0}\right)^{\alpha}}\right)}_{\substack{improved\ \ factor \\ allocation}} + \underbrace{\ln\left(\frac{H_0^{1-\alpha}\cdot K_0^{\alpha}}{H_{c,0}^{1-\alpha}\cdot K_{c,0}^{\alpha}}\right)^{\mu-1}}_{\substack{increased \\ market\ \ size}} \geq 0$$

where $Y^{I}_{c,0}$ is the actual income of the inhabitants of region c at date t=0, and $Y^{A}_{c,0}$ is the income they would have had at date t=0 if globalization had not taken place. Since each of the terms in Equation (28) is non-negative, the first result we obtain is that the overall impact or static gains from economic integration are non-negative as well.

These gains can be decomposed into three sources corresponding to each of the terms of Equation (28). The first one shows the growth of income that results

[19] To derive this expression I have assumed a zero cross-industry correlation between α_i and μ_i, i.e. υ=0. This parameter restriction is useful because it allows us to unambiguously disentangle the "increased-market-size" and "improved-factor-allocation" effects.

from moving industries from low to high productivity locations. This term would vanish if region c had the highest productivity in all industries. The second term shows the growth of income that results from relocating factors away from those regions and/or industries in which they were abundant in autarky into those in which they were scarce. This term would vanish if region c had world average factor proportions. The third term shows the growth in income that is due to an increase in market size that allows industries to support a higher degree of specialization. This term would vanish if the size of region c were arbitrarily large with respect to the rest of the world. An implication of Equation (28) is that the static gains from economic integration are greater for regions with low productivity, extreme factor proportions and modest amounts of physical and human capital.

If coupled with an appropriate transfer scheme, globalization leads to a Pareto improvement in the world economy. Equation (28) shows that, with the same production factors, the integrated economy generates more output than the world of autarky. It is therefore possible to implement a transfer scheme that keeps constant the income of all current and future young and gives more income to all current and future old. Under this transfer scheme, investment and the stock of physical capital would be unaffected by economic integration. But the production and consumption of all generations born at date t=0 or later would increase. Of course, there exist many alternative transfer schemes that ensure that globalization benefits all. Moreover, since each region gains from trade there exist Pareto-improving transfer schemes that can be implemented without the need for inter-regional transfers. That is, ensuring that globalization generates a Pareto improvement does not require compensation from one region to another.

How “large” the transfer scheme must be to ensure that economic integration leads to a Pareto improvement? The answer is “not much” if most of the gains from economic integration come from higher productivity and increased market size. The reason is that in this case all factors share in the gains from integration. The required transfer scheme could be “substantial” if the gains from integration come mostly from

improved factor allocation. This is because within each region the owners of the abundant factor obtain more than proportional gains from integration while the owners of the region's scarce factor might have losses. In this case, implementing a Pareto improvement requires a transfer from the former to the latter.

Without a transfer scheme, it is relatively straightforward to trace the dynamic effects of economic integration. Assume for simplicity that the world contains many symmetric regions so that before integration all of them had the same law of motion. The top panel of Figure 7 shows the effects of economic integration in the "deterministic" world when diminishing returns are strong and market size effects are weak. Economic integration raises the steady state stock of physical capital and sets up a period of high growth that eventually ends. It is straightforward to see that the effects would be similar in the "stochastic" world, with economic integration permanently raising the steady state interval. Using the jargon of growth theory, if $\mu\cdot\alpha+\upsilon<1$ economic integration has level effects on income. The bottom panel of Figure 7 shows the opposite case in which diminishing returns are weak and market size effects are strong. In this case, economic integration shifts down the steady state value, increasing the growth rate permanently. Once again, it is straightforward to see that the effects would be similar in the "stochastic" world, with trade shifting the threshold interval to the left. Using again the jargon of growth theory, if $\mu\cdot\alpha+\upsilon\geq1$ integration has growth effects on income.

It is tempting now to revisit our earlier account of the history of the world economy since the 1500s, and propose an alternative version which is also based on the "stochastic" world with $\mu\cdot\alpha+\upsilon>1$. It goes as follows: for centuries, the world economy consisted of a collection of autarkic regions that were too small to sustain economic growth. Located within the threshold interval, these regions were subject to periodic expansions and contractions with virtually zero average growth. Once again, this is consistent with Maddison's calculation that the world economy grew only about eighteen percent from 1500 to 1820. The Industrial Revolution occurs when a series of reductions in trade costs between some British regions raised their

combined size above the threshold interval and set them on the path of accelerating growth. As time went on, more and more regions joined the initial core and the Industrial Revolution spread throughout Britain and moved into France, Germany and beyond. It is therefore a reduction of trade costs and the progressive extension of markets that made possible sustained growth and allowed the world economy to grow more than seven hundred and fifty percent from 1820 to 1998. This might also explain why this growth in world income was accompanied by an even higher growth in world trade.[20]

This view of the development process is also broadly consistent with the general observations about inequality between center and periphery discussed in the introduction. Regions that join the integrated economy (the "center") become rich and take off into steady growth. Regions that do not join the integrated economy (the "periphery") are left behind, technologically backward and capital poor. As more and more regions enter the integrated economy, those that are left behind become relatively poorer and world inequality increases. Eventually all regions will enter the integrated economy and world inequality will decline. Therefore, this model generates an inverted U-shape or Kuznets curve, with world inequality rising in the first stages of world development and declining later. Pritchett [1997], Bourguignon and Morrison [2002] and others have shown that world inequality has increased from 1820 to now. It remains to be seen if this inequality will decline in the future.

This stylized model also illustrates some of the conflicts that globalization might create. It follows from Equation (28) that the gains from trade are large for regions whose factor proportions are far from the world average. Ceteris paribus, this means that regions in the center would like that new entrants into the integrated economy to move the world average factor proportions away from them. In fact, unless productivity and market size effects are substantial, the entry of a large region creates losses to other regions with similar factor proportions. This implies, for instance, that the Chinese process of economic integration should be seen with

[20] The word "might" reflects the earlier observation that regional production and therefore trade is indeterminate in the integrated economy.

some concern in countries with similar factor proportions such as Mexico and Indonesia, but with hope in the European Union or the United States.

This view of globalization and growth leads to a powerful prescription for economic development: open up and integrate into the world economy. I believe this is a fundamentally sound policy prescription, and history is largely consistent with it. But there are a number of important qualifications that this stylized model cannot capture. Integrating into the world economy is not an “all-or-nothing” type of affair in which regions move overnight from autarky to complete integration. The process of economic integration is slow and full of treacherous steps. Obtaining general prescriptions for development in a world of imperfect integration has proved to be a much more challenging task. I shall come back to this important point later, but we must first introduce trade frictions into the story.

2. Specialization, trade and diminishing returns

Let us revise our model of globalization. As in section 1.3, assume that at date t=0 the costs of transporting goods across regions suddenly fall from prohibitive to negligible. Unlike section 1.3, assume now that the costs of transporting factors across regions remain prohibitive after date t=0. An implication of this setup is that globalization equalizes goods prices across regions, but it does not necessarily equalize factor prices. This particular view of globalization has a longstanding tradition in trade theory and the goal of this section is to analyze it.

Assuming that human capital is immobile internationally is somewhat dubious, as there are some well-known examples of large contingents of people working overseas. But most of the results discussed here would go through with only minor changes under the weaker and reasonably realistic assumption that international

flows of people are quantity constrained, although not necessarily at zero.[21] Assuming that physical capital is immobile is appropriate for buildings and structures and, probably, not too unreasonable for the most important types of machinery and equipment. Moreover, assuming that existing physical capital cannot be transported does not preclude physical capital to effectively "move" across regions over time, as it declines in some regions through depreciation and increases in others through investment.[22]

If physical capital is immobile, pieces of capital located in different regions might offer different return distributions. This opens up a role for financial markets. Although the old and the impatient young still have no incentive to trade securities, the patient young now have a motive. Those that are located in regions where physical capital offers an attractive distribution of returns want to sell securities and use the proceeds to finance additional purchases of domestic physical capital. Those patient young that are located in regions where physical capital offers an unattractive distribution of returns want to buy securities and reduce their holdings of domestic physical capital. And, regardless of their location, the patient young want to buy and sell securities in order to share regional risks. Thus, the immobility of physical capital creates a potentially important role for international financial markets: the geographical reallocation of investments and production risks.

Despite this, I will not let international financial markets play this role. This failure of financial markets could be due to technological motives or informational problems of various sorts. But I prefer instead to think of it as being caused by lack of incentives to enforce international contracts. In the integrated economy,

[21] Of course, this becomes a weak or empty excuse if quantity constraints respond to economic incentives in a systematic way. See Lundborg and Segerstrom [2002] and Ortega [2004] for models in which this happens.

[22] Remember that we have assumed that physical capital depreciates in one generation. Therefore, assuming physical capital is immobile only means that it is not possible at date t to move around the stock of physical capital created and deployed at date t-1, and that is being used for production at date t. But it is certainly possible to choose where to deploy the new stock of physical capital created at date t that will be used for production at date t+1. The effects of physical capital immobility would be more severe quantitatively with a slower rate of depreciation. Note also that immobility matters only because physical capital is irreversible or putty-clay. In fact, it would be logically inconsistent to assume that physical capital is immobile if it could be converted back into mobile goods.

individuals could enter into contracts that specify exchanges of various quantities of the different goods to be delivered at various dates and/or states of nature. It is standard convention to refer to the signing of contracts that involve only contemporaneous deliveries as “goods” trade, while the signing of contracts that involve future (and perhaps state contingent) deliveries is usually referred to as “asset” trade. Both types of trade require sufficiently low costs of transporting goods. But asset trade also requires that the signing parties credibly commit to fulfill their future contractual obligations. The domestic court system punishes those that violate contracts, thus creating the credibility or trust that serves as the foundation for domestic financial markets. But there is no international court system that endows sovereigns with the same sort of credibility, and this hampers international financial markets. I assume next this problem is so severe that it precludes all asset trade.

Unlike the integrated economy, in the world analyzed in this section each region’s total production, spending and capital stock are always determined. Since trade balances and current accounts are zero, the income of each region equals the value of both its production and spending, i.e. $Y_{c,t}=Q_{c,t}=E_{c,t}$. Since the only vehicle for savings available to the young is physical capital, analogues to Equations (1)-(2) apply to each region. We can therefore write regional incomes and the laws of motion of regional capital stocks as follows:

(29) $Y_{c,t} = w_{c,t} \cdot H_{c,t} + r_{c,t} \cdot K_{c,t}$ for all $c \in C$

(30) $K_{c,t+1} = S_{c,t} \cdot w_{c,t} \cdot H_{c,t}$ for all $c \in C$

These Equations apply to all the models of this section, including the world of autarky before globalization. Therefore, a complete analysis of the world income distribution and its evolution requires us to determine the cross-section of factor prices, i.e. $w_{c,t}$ and $r_{c,t}$ as a function of the state of the world economy. The latter consists of the savings, factor endowments and industry productivities of all regions

of the world, i.e. $S_{c,t}$, $H_{c,t}$, $K_{c,t}$ and $Z_{c,it}$ for all $i \in I$ and all $c \in C$; plus the date, since trade in goods is only possible if $t \geq 0$.

The rest of this section is organized as follows. Sub-section 2.1 studies further the world of autarky, while the rest of the section studies the world after globalization. In sub-section 2.2, we explore a world in which frictions to factor mobility and asset trade are not binding after globalization. Sub-section 2.3 provides a formal description of the model. Sub-sections 2.4 and 2.5 examine worlds where frictions to factor mobility and asset trade remain binding after globalization.

2.1 Economic growth in autarky

The analysis of the effects of globalization starts in the world of autarky. As explained in section 1.3, before globalization each region is a smaller and less efficient version of the integrated economy and factor prices can be written as:[23]

(31) $$w_{c,t} = (1-\alpha) \cdot A_{c,t} \cdot H_{c,t}^{\mu \cdot (1-\alpha) - \upsilon - 1} \cdot K_{c,t}^{\mu \cdot \alpha + \upsilon} \qquad \text{for all } c \in C$$

(32) $$r_{c,t} = \alpha \cdot A_{c,t} \cdot H_{c,t}^{\mu \cdot (1-\alpha) - \upsilon} \cdot K_{c,t}^{\mu \cdot \alpha + \upsilon - 1} \qquad \text{for all } c \in C$$

Equations (31)-(32) describe the cross-section of factor prices. Holding constant factor endowments, regions with higher than average industry productivities have higher than average factor prices. Holding constant industry productivities, the relationship between factor prices and factor endowments depends on two familiar forces: diminishing returns and market size. For a given set of industry technologies, an increase in one factor makes this factor relatively more abundant, lowering its price and raising the price of the other factor. But an increase in one factor also raises income and demand in all industries, improving industry technologies and

[23] These Equations follow from Equation (24) and the observation that the shares of human capital and physical capital are 1-α and α.

raising the prices of both factors. Equations (31)-(32) put these two effects together. Hence, regions with higher-than-average human capital have higher-than-average rental rates for all parameter values, and also higher-than-average wages if $\mu\cdot(1-\alpha)-\upsilon>1$. Similarly, regions with higher-than-average physical capital have higher-than-average wages for all parameter values, and also higher-than-average rental rates if $\mu\cdot\alpha+\upsilon>1$.

It follows from Equations (29)-(32) that, before globalization, we can write regional incomes and capital stocks as follows:[24]

(33) $Y_{c,t} = A_{c,t} \cdot H_{c,t}^{\mu\cdot(1-\alpha)-\upsilon} \cdot K_{c,t}^{\mu\cdot\alpha+\upsilon}$ for all $c \in C$

(34) $K_{c,t+1} = s_{c,t} \cdot A_{c,t} \cdot H_{c,t}^{\mu\cdot(1-\alpha)-\upsilon} \cdot K_{c,t}^{\mu\cdot\alpha+\upsilon}$ for all $c \in C$

Equation (33) shows the income of regions, and it can be used to determine the relative contribution of factor endowments and productivity to income differences. For instance, assume income is λ times higher than average in a given region. It could be that in this region human capital is $\lambda^{1/(\mu\cdot(1-\alpha)-\upsilon)}$ higher than average or that physical capital is $\lambda^{1/(\mu\cdot\alpha+\upsilon)}$ higher than average. It could also be that the region's productivity in industry i is $\lambda^{1/\sigma_i\cdot\mu_i}$ times higher than average.[25] Naturally, it could also be any combination of these factors.

Equation (34) is the law of motion of the capital stocks and can be used to analyze the dynamic response to a region-specific shock to savings, human capital and/or industry productivity. Positive (and permanent) shocks to any of these variables raise the region's capital stock and income. As Equation (34) shows, these shocks have growth effects if $\alpha\cdot\mu+\upsilon\geq1$, but only have level effects if $\alpha\cdot\mu+\upsilon<1$. Regardless of the case, the effects of these shocks never spill over to other regions.

[24] These Equations are identical to Equations (24)-(25) and have been reproduced here only for convenience.
[25] Here industry productivity means $Z_{c,it}$, and not $A_{c,it}$.

Assume the joint distribution of savings, human capital and industry productivities is stationary. Then, Equations (33)-(34) imply a strong connection between the cross-sectional and time-series properties of the growth process.. If diminishing returns are strong and market size effects are weak, i.e. if $\mu\cdot\alpha+\upsilon<1$, world average income ($Y_t$) and its regional distribution ($Y_{c,t}/Y_t$) are both stationary. If instead diminishing returns are weak and market size effects are strong, i.e. $\mu\cdot\alpha+\upsilon>1$, world average income and its regional distribution are both non-stationary. This result provides a tight link between the long run properties of the growth process and the stability of the world income distribution. A weaker version of this result assumes that the world productivity frontier (A_t) is non-stationary but regional productivity gaps ($A_{c,t}/A_t$) are stationary. Under this assumption, world average income is non-stationary even if diminishing returns are strong and market size effects are weak.

It is commonplace among growth theorists to interpret cross-country data from the vantage point of the autarky model.[26] One influential example is the work of Mankiw, Romer and Weil [1992]. They combined Equations (33)-(34) to obtain an Equation relating income to savings, human capital, country productivity, and lagged income; and estimated it using data for a large cross-section of countries. They interpreted the residuals of this regression as measuring differences in country productivities and measurement error, and concluded that differences in savings and human capital explain (in a statistical sense) about 80 percent of the cross-country variation in income. Their procedure imposed the restriction $\mu=1$ (and therefore $\upsilon=0$) and yielded an estimate of α of about two thirds. Hall and Jones [1999] and Klenow and Rodríguez-Clare [1997] interpreted this high estimate of α as a signal that the regression was miss-specified. Their argument was that savings, human capital and productivity were positively correlated and the omission of productivity from the

[26] Unfortunately, the absence of direct and reliable measures of productivity precludes carrying out formal tests of the theory. The most popular empirical response to this problem has been to simply assume the theory is correct and use available data to make inferences about the determinants of the world income distribution and its evolution.

regression biased upwards the estimate of α. These authors used Equations (33)-(34) to calibrate country productivities keeping the assumption that μ=1, but instead imposing a value of α of about one third.[27] With these productivities at hand, they found that about two thirds of the variation in incomes reflects variation in productivity, and only one third can be attributed to cross-country variation in savings and human capital.

Another influential example of the use of the autarky model to interpret available data is Barro [1991] who found that, after controlling for human capital and saving rates, poor countries tend to grow faster than rich ones. This finding has been labeled "conditional convergence" since it implies that, if two countries have the same country characteristics, they converge to the same level of income.[28] If Equations (33)-(34) provide a good description of the real world, observing conditional convergence is akin to finding that $\mu \cdot \alpha + \upsilon < 1$.[29] Many have therefore interpreted the conditional convergence finding as evidence that diminishing returns are strong relative to market size effects.

These inferences about the nature of the growth process heavily rely on Equations (33)-(34), and these Equations have been derived from a theoretical model that assumes that all regions of the world live in autarky. This assumption is obviously unrealistic. Is it also crucial? And if so, what alternative assumption would be reasonably realistic? I next turn to these questions. But the script should not be surprising. Globalization (as described at the beginning of this section) has profound effects on the world income distribution and its evolution. The newfound ability of regions to specialize and trade alters, sometimes quite dramatically, the effects of

[27] This value corresponds to the share of capital in income in national accounts. This sort of calibration exercise is known as development accounting. Caselli's chapter in this volume is the definitive source on this topic.

[28] As Barro himself emphasized, this does not mean that per capita incomes tend to converge unconditionally since countries with high initial incomes also tend to have good country characteristics. There is a large number of papers that try to determine whether there is conditional convergence and measure how fast it takes place. See, for instance, Knight, Loayza and Villanueva [1993] and Caselli, Esquivel and Lefort [1996].

[29] An additional maintained assumption of this line of research is that savings, human capital and productivity are jointly stationary.

factor endowments and industry productivities on factor prices. This is most clearly illustrated in subsection 2.2, which depicts a world in which goods trade allows the world economy to replicate the prices and allocations of the integrated economy. Of course, this is not a general feature of goods trade. Subsection 2.3 prepares the ground for the analysis of worlds where economic integration is imperfect and factor prices vary across regions. This analysis is then performed in subsections 2.4 and 2.5.

2.2 Factor price equalization

A good starting point for the analysis of the world economy after globalization is to ask whether restricting factor mobility matters at all. Somewhat surprisingly, the answer is "perhaps not". As Paul Samuelson [1948, 1949] showed more than half a century ago, goods trade might be all that is needed to ensure global efficiency. When this happens, we say that the equalization of goods prices leads to the equalization of factor prices. I shall describe Samuelson's result and its implications step by step, so as to develop intuition.[30]

Consider the set of all possible partitions of the world factor endowments at date t, H_t and K_t, among the different regions of the world or, for short, the set of all possible factor distributions. This set is formally defined as follows:

$$(35) \qquad D_t \equiv \left\{ (H_{c,t}, K_{c,t}) \text{ for all } c \in C \;\middle|\; H_{c,t} \geq 0,\ K_{c,t} \geq 0 \ \text{ s.t. } \sum_{c \in C} H_{c,t} = H_t \text{ and } \sum_{c \in C} K_{c,t} = K_t \right\}$$

Define FPE_t as the subset of D_t for which the world economy replicates the prices and allocations of the integrated economy. To construct FPE_t, fix $d_t \in D_t$ and consider the integrated economy prices and quantities. At these prices, consumers

[30] The analysis here follows a long tradition in international trade. See Norman and Dixit [1989], Helpman and Krugman [1985] and Davis [1995].

are willing to purchase the integrated economy quantities of the different goods and also have enough income to do so. At these prices, producers located in regions with the highest industry productivities are willing to produce the integrated economy quantities of the different goods using the integrated economy quantities of factors. If these producers can find these quantities of factors in their regions, the integrated economy prices and quantities are in fact the equilibrium ones and we say that $d_t \in FPE_t$. Otherwise, the integrated economy prices and quantities cannot be the equilibrium ones and we say that $d_t \notin FPE_t$. Therefore, the set FPE_t can be formally defined as follows:

$$
\begin{aligned}
FPE_t \equiv \Bigg\{ d_t \in D_t \;\Big|\; & \exists x_{c,it}(m) \geq 0,\ x^F_{c,it} \geq 0 \text{ with } \sum_{c \in C} x_{c,it}(m) = 1,\ \sum_{c \in C} x^F_{c,it} = 1 \text{ and} \\
& x_{c,it} = (1-\beta_i) \cdot x^F_{c,it} + \frac{\beta_i}{M_{it}} \cdot \int_0^{M_{it}} x_{c,it}(m) \cdot dm; \text{ such that :} \\
& \text{(R1) } x_{c,it} = 0 \text{ if } Z_{c,it} < \max_{c \in C} \{Z_{it}\}; \\
& \text{(R2) } H_{c,t} = \sum_{i \in I} x_{c,it} \cdot H_{it} \text{ and } K_{c,t} = \sum_{i \in I} x_{c,it} \cdot K_{it}; \text{ and} \\
& \text{(R3) } x_{c,it}(m) \in \{0,1\} \text{ for all } m \in [0, M_{it}] \text{ and } i \in I \Bigg\}
\end{aligned}
\tag{36}
$$

where M_{it}, H_{it} and K_{it} are defined in Equations (17), (19) and (20). To understand this definition, interpret $x_{c,it}$ as the share of the world production of industry i located in region c at date t; and note that this share includes the production of intermediate inputs, $x_{c,it}(m)$, and final goods, $x^F_{c,it}$. Definition (36) then says that $d_t \in FPE_t$ if it is possible to achieve full employment of human and physical capital in all regions producing only in those regions with the highest productivity (requirement R1), using the same factor proportions as in the integrated economy (requirement R2), and without incurring the fixed cost of production more than once (requirement R3). The set FPE_t is never empty since the factor distribution that applies in the integrated economy always belongs to it. In fact, the set FPE_t consists of all the factor

distributions that are equilibria of the integrated economy. The larger is the size of the indeterminacy in the geographical distribution of production and factors of the integrated economy, the larger is the size of FPE_t.

The patterns of production and trade that support factor price equalization after globalization are easy to state and quite intuitive:

1. *In regions where human (physical) capital is relatively abundant, production shifts towards industries that, on average, use human (physical) capital intensively. Excess production in these industries is converted into exports that finance imports of industries that use physical (human) capital intensively.*

Example 2.1.1: Consider a world economy with H- and K-industries, such that $I^H \cup I^K = I$ and $I^H \cap I^K = \varnothing$. Assume $\alpha_i=\alpha_H$ if $i \in I^H$, and $\alpha_i=\alpha_K$ if $i \in I^K$; and $\alpha_H<\alpha_K$; and $\beta_i=0$ for all $i \in I$. All regions have the same industry productivities, but A-regions have a higher ratio of human to physical capital than B-regions. Factor price equalization is possible if the differences in factor proportions between A- and B-regions are not too large relative to the differences in factor proportions between H- and K-industries. Figure 8 shows the geometry of this example. Since all regions have the same factor costs, industries use the same factor proportions in all regions. A-regions contain a more than proportional fraction of the integrated economy's H-industry, and a less than proportional fraction of the K-industry. The opposite happens in B-regions. This is how specialization and trade ensure that in this world economy factor endowments are used efficiently.

2. *In industries where a region's productivity is less than the world's highest, production falls to zero and domestic spending shifts towards imports. To finance*

the latter, production expands in industries in which the region has the highest possible productivity and the excess production is exported abroad.

Example 2.1.2: Consider a world economy with H- and K-industries, such that $I^H \cup I^K = I$ and $I^H \cap I^K = \varnothing$. Assume $\alpha_i=\alpha_H$ if $i \in I^H$, and $\alpha_i=\alpha_K$ if $i \in I^K$; and $\alpha_H<\alpha_K$; and $\beta_i=0$ for all $i \in I$. Within each type there are "advanced" and "backward" industries. A-regions have the highest possible productivity in all industries, regardless of whether they are "advanced" or "backward". B-regions have the highest possible productivity only in "backward" industries. Factor price equalization is possible if the combined factor endowments of A-regions are large enough and the subset of "advanced" industries is not too large. Figure 9 shows the geometry of this example. Since all regions have the same factor costs, only producers located in regions with the highest productivity can survive international competition. A-regions produce the integrated economy quantities of "advanced" goods and a fraction of the integrated economy quantities of "backward" goods. B-regions produce the remaining quantities of "backward" goods. This is how specialization and trade ensure that in this world economy production takes place only where industry productivities are higher.

3. *Within each industry, only one region produces each input variety and exports it to all other regions. If an industry is split among various regions, there is likely to be two-way trade within the same industry*.[31]

Example 2.1.3: Consider any of the world economies of the previous examples, but assume now that $\beta_i=1$ for all $i \in I$. Assume $d_t \in FPE_t$. Since the fixed costs of

[31] I say "likely to be" because a region might produce the final good for domestic use, and import the necessary input varieties. It is usual in trade models to set $\beta_i=1$ and then drop the "likely to be" from the statement.

producing inputs contain the cost of building a specialized production plant, all input producers choose to concentrate their production in one region in order not to duplicate these costs. Therefore, each region produces a disjoint set of input varieties. This is how specialization and trade allow the world economy to exploit increasing returns to scale and therefore benefit from a larger market size.

By adopting these patterns of specialization and trade, the world economy is able to reap all the benefits of economic integration without any factor movements. Using the jargon of trade theory, goods trade is a "perfect substitute" for factor movements if $d_t \in FPE_t$. When this is the case, factor prices are given by:

(37) $w_{c,t} = (1-\alpha) \cdot A_t \cdot H_t^{\mu\cdot(1-\alpha)-\upsilon-1} \cdot K_t^{\mu\cdot\alpha+\upsilon}$ for all $c \in C$

(38) $r_{c,t} = \alpha \cdot A_t \cdot H_t^{\mu\cdot(1-\alpha)-\upsilon} \cdot K_t^{\mu\cdot\alpha+\upsilon-1}$ for all $c \in C$

The world economy is able to operate at the same level of efficiency as the integrated economy despite the immobility of factors. Equations (31)-(32) showed that, before globalization, cross-regional differences in factor proportions and industry productivities lead to differences in the way industries operate (i.e. their factor proportions and productivity) and also in the size of their markets. Regions with a high ratio of human to physical capital have high wage-rental ratios. Regions with high industry productivities and abundant human and physical capital have high factor prices. But Equations (37)-(38) show that, after globalization (and if $d_t \in FPE_t$), cross-regional differences in factor proportions and industry productivities neither change the way industries operate, nor do they affect the size of their markets. Goods trade allows regions to absorb their differences in factor endowments and industry productivities by specializing in those industries that use their abundant factors and have the highest possible productivity, without the need for having different factor prices. Goods trade also eliminates the effects of regional size on factor prices by creating global markets.

These observations have important implications for the world income distribution and, consequently, for any attempt to determine the relative contribution of factor endowments and productivity to income differences. Substituting Equations (37)-(38) into Equation (29), we find that:

$$(39) \qquad Y_t = A_t \cdot H_t^{\mu\cdot(1-\alpha)-\upsilon} \cdot K_t^{\mu\cdot\alpha+\upsilon}$$

$$(40) \qquad \frac{Y_{c,t}}{Y_t} = (1-\alpha)\cdot\frac{H_{c,t}}{H_t} + \alpha\cdot\frac{K_{c,t}}{K_t} \qquad \text{for all } c\in C$$

A comparison between these Equations and Equation (33) shows that the relative contribution of factor endowments and productivity to income differences is fundamentally affected by globalization. Equation (33) differs from Equations (39)-(40) in three important respects: the elasticity of substitution between domestic human and physical capital is one in Equation (33) but infinite in Equations (39)-(40); domestic productivity appears in Equation (34) but not in Equations (39)-(40); and income is homogeneous of degree μ on domestic factor endowments in Equation (34) but only of degree one in Equation (39)-(40). Each of these differences echoes a different aspect of globalization, and I shall discuss them in turn.

Globalization raises the elasticity of substitution between human and physical capital from one to infinity because structural transformation (a shift towards industries that use the abundant factor) replaces factor deepening (forcing industries to use more of the locally abundant factor) as a mechanism to absorb differences in factor proportions. Assume a region has a ratio of human to physical capital λ times higher than average. Before globalization, each of its industries is forced to operate with a ratio of human to physical capital that is λ times average, and this requires a wage-rental ratio that is λ^{-1} times average. After globalization, the region simply shifts its production towards industries that are human-capital intensive, keeping the ratio of human to physical capital of its industries constant. This does not require changes in the wage-rental ratio.

Globalization eliminates differences in industry productivities as a source of income differences because structural transformation (a shift towards industries that have high productivity) also replaces productivity deepening (forcing low-productivity industries to produce) as a mechanism to absorb differences in industry productivities. Assume now that a region has average factor endowments but higher than average industry productivities. For instance, the region's productivity is λ times higher than the rest of the world in a subset of industries of combined size σ, and equal to the rest of the world in the remaining ones. Before globalization, this productivity advantage allows the region to produce $\lambda^{\sigma \cdot \mu}$ output than average with the same factors, holding constant technology. After globalization, the region takes over all world production of those industries in which its productivity is higher and scales back the rest of its industries. This allows the rest of the world to take full advantage of the region's high productivity and catch up with it in terms of income (even though not in productivity).

Globalization reduces the effects of factor endowments on relative incomes because it converts regional markets into global ones. Assume now that a region has average industry productivities, but its human and physical capitals are both λ times above average. Before globalization, the region's higher factor endowments allow it to produce more output than the average region. This effect is further reinforced because the region's larger market size allows it to have a better technology than average. Therefore, in autarky the region's income is λ^{μ} times higher than the world's average. After globalization, this additional market size effect disappears since the relevant market is the world market and this is the same for all regions. Therefore, after globalization the region's income is only λ times higher than the world average income.

Globalization also influences the dynamics of the world economy. Assume $d_t \in FPE_t$ for all t, then it follows from Equation (30) and (37)-(38) that:

(41) $$K_{t+1} = s_t \cdot A_t \cdot H_t^{\mu\cdot(1-\alpha)-\upsilon} \cdot K_t^{\mu\cdot\alpha+\upsilon}$$

(42) $$\frac{K_{c,t+1}}{K_{t+1}} = \frac{S_{c,t} \cdot H_{c,t}}{S_t \cdot H_t} \qquad \text{for all } c \in C$$

A comparison between these Equations and Equation (34) shows how globalization affects the dynamic responses to region-specific shocks. After globalization, positive (and permanent) shocks to savings and human capital still raise a region's capital stock and income. But now the effects of these shocks spill over to other regions. Shocks to productivity can only affect a region's income if they push outward the world productivity frontier. And, in this case, all countries equally benefit.[32]

Another important implication of Equations (39)-(42) is that globalization breaks down the connection between the long run properties of the growth process and the stability of the world income distribution.[33] Assume again that the joint distribution of savings, human capital and productivities is stationary. Then, Equation (41) shows that it still is the relative strength of diminishing returns and market size effects that determines whether world average income is stationary or not. But Equation (42) shows that now the world distribution of capital stocks is stationary regardless of parameter values. The same applies to the world income distribution (see Equation (40)). Therefore, all regions share a common growth rate in the long run. The reason is simple: physical capital accumulation in high-savings and high-human capital regions is absorbed by increased production in industries that use physical capital intensively, and this lowers the prices of these industries and increases the prices of industries that use human capital intensively. This increases wages and savings in low-savings and low-human capital regions. In a nutshell,

[32] See Ventura [1997] and Atkeson and Kehoe [2000] for analyses of shocks to small open economies in the factor-price-equalization world.
[33] Ventura [1997] provides a dramatic example of this by constructing a world in which time-series convergence to a steady state is associated with cross-sectional divergence and vice versa.

movements in goods prices positively transmit growth across regions and ensure the stability of the world income distribution. [34]

The main feature of the factor-price-equalization world is that diminishing returns and market size effects are global and not local. This observation has important implications for growth theory. Explanations for why the world grows faster today than in the past should feature diminishing returns and market size effects in the lead role, and relegate savings and human capital to a secondary one. But explanations of why some countries grow faster than others should do exactly the opposite, giving the lead role to savings and human capital and relegating diminishing returns and market size effects to a secondary role. A distinctive feature of the integrated economy is therefore a sharp disconnect between the determinants of average or long run growth and the determinants the dispersion or the cross-section of growth rates.[35]

The factor-price-equalization world neatly illustrates the potential effects of trade on the world income distribution and its dynamics, and it shows why and how goods trade can be a perfect substitute for factor movements. But the real world has not achieved yet the degree of economic integration that this model implies. One does not need sophisticated econometrics to conclude that wages vary substantially around the world. It is less obvious but probably true as well that rental rates also vary substantially around the world. These differences in factor prices indicate that regional differences in factor endowments and/or industry productivities are so large that goods trade cannot make up for factor immobility.

---

[34] As a general proposition, it is not necessary that trade leads to the stability of the world income distribution. In fact, the study of the stability of world income distribution has received considerable attention recently. While Acemoglu and Ventura [2002] rely on specialization to generate a stable world income distribution, Deardorff [2001] presents a model in which mere differences in initial endowments create persistent difference in world income and "club convergence". Krugman [1987] and Howitt [2000] rely on endogenous technology change to generate such effects. See Brezis, Krugman, and Tsiddon [1993] for a model of human capital accumulation that explains leapfrogging in the international income distribution.

[35] One implication of this is that Barro's conditional convergence finding cannot be used to determine whether diminishing returns are weak or strong relative to market size effects. See Ventura [1997].

What trade always does is to create a global market in which only the most competitive producers of the world can survive. Trade forces high-cost industries to close down and offers low-cost industries the opportunity to grow. If $d_t \in FPE_t$, all regions contain enough of these low-cost industries to employ all of their factors at common or equalized factor prices. But this need not be always the case. If $d_t \notin FPE_t$, regions with low industry productivities and sizeable factor endowments are forced to offer cheap factors to compete, while regions with high industry productivities and small factor endowments are able to enjoy expensive factors. These price differences indicate that factors are not deployed where they should and the world economy does not operate efficiently. To study the origins and effects of these world inefficiencies, it is necessary first to review some formal aspects of the model after globalization.

2.3 Formal aspects of the model

As mentioned already, in the absence of asset trade analogues of Equations (1) and (2) apply now to each region of the world economy. A regional analogue to Equation (3) also applies since it is a direct implication of our Cobb-Douglas assumption for the consumption and investment composites. Since all regions share spending patterns and face the same goods prices, the price of consumption and investment is the same for all. We keep this common price as the numeraire and, as a result, Equation (4) also applies. Equations (5)-(6) describing technology apply to all regions, with the corresponding factor prices and industry productivities.

After globalization, Equations (7)-(10) describing pricing policies, input demands and the free-entry condition apply only to those regions that host the lowest-cost producers of the world. The rest cannot compete in global markets. To formalize this notion, define the following sets of industries:

$$(43) \qquad I_{c,t} \equiv \left\{ i \in I \;\middle|\; c \in \arg\min_{c' \in C} \left\{ \frac{1}{Z_{c',it}} \cdot \left(\frac{w_{c',t}}{1-\alpha_i} \right)^{1-\alpha_i} \cdot \left(\frac{r_{c',t}}{\alpha_i} \right)^{\alpha_i} \right\} \right\} \qquad \text{for all } c \in C$$

An industry belongs to $I_{c,t}$ if and only if producers located in region c are capable of competing internationally in this industry at date t.[36] Note that a region can be competitive in a given industry because it offers high productivity or a cheap combination of factor prices. The main implication of goods trade is that industries do not locate in regions where they are not competitive:

$$(44) \qquad Q_{c,it} = 0 \qquad \text{if } i \notin I_{c,t} \qquad \text{for all } i \in I \text{ and } c \in C$$

Since goods markets are integrated, Equation (11) describing market clearing in global goods markets still applies. But now Equations (12)-(13) describing market clearing in global factor markets must be replaced by analogue conditions imposing market clearing in each regional factor market:

$$(45) \qquad H_{c,t} = \sum_{i \in I} H_{c,it} \qquad \text{with} \quad H_{c,it} = \frac{\partial B_{c,it}}{\partial w_{c,t}} + \int_0^{M_{it}} \frac{\partial b_{c,it}(m)}{\partial w_{c,t}} \cdot dm \qquad \text{for all } c \in C$$

$$(46) \qquad K_{c,t} = \sum_{i \in I} K_{c,it} \qquad \text{with} \quad K_{c,it} = \frac{\partial B_{c,it}}{\partial r_{c,t}} + \int_0^{M_{it}} \frac{\partial b_{c,it}(m)}{\partial r_{c,t}} \cdot dm \qquad \text{for all } c \in C$$

Equations (45)-(46) state that the regional supplies of labor and capital must equal their regional demands. The latter are the sum of their industry demands, and these are calculated by applying Shephard's lemma to Equations (5) and (6).

[36] This follows directly from the cost functions in Equations (5)-(6) and the observation that all producers in the world face the same world demand.

This completes the formal description of the model. For any admissible set of capital stocks, i.e. $K_{c,0}$ for all $c \in C$, and sequences for the vectors of savings, human capital and industry productivities, i.e. $S_{c,t}$, $H_{c,t}$, and $A_{c,it}$ for all $c \in C$ and for all $i \in I$, an equilibrium of the world economy after globalization consists of sequences of prices and quantities such that the equations listed above hold at all dates and states of nature. Although there might be multiple geographical patterns of production and trade that are consistent with world equilibrium, the assumptions made ensure that prices and world aggregates are uniquely determined.[37]

We are ready now to re-examine the effects of globalization on factor prices and the world income distribution. We have already found that, if $d_t \in FPE_t$, globalization eliminates all regional differences in factor prices and permits the world economy to operate at the same level of efficiency as the integrated economy. In this case, global market forces are strong enough to ensure that diminishing returns and market size effects have a global rather than a regional scope. This is no longer the case if $d_t \notin FPE_t$ since globalization cannot eliminate all regional differences in factor prices. These factor price differences reflect inefficiencies of various sorts in the world economy.

Efficiency requires that factor usage within an industry be the same across regions. This is a direct implication of assuming diminishing returns to each factor in production. The problem, of course, is that regional factor proportions vary. Structural transformation allows regions to accommodate all or part of their differences in factor proportions without factor deepening. If there are enough industries that use different factor proportions, factor prices are equalized across regions. If there are not enough industries that use different factor proportions, regions must lower the price of their abundant factor and raise the price of their

[37] Despite the indeterminacy in trade patterns, the trade theorist will immediately recognize that, if $\beta_i=1$ for all $i \in I$, the volume of trade is determined and the popular gravity equation applies to this world economy.

scarce one to attract enough firms to employ their factor endowments. In this case, industries in different regions use different factor proportions and the world economy is inefficient. Subsection 2.4 studies the properties of the growth process in this situation.

Efficiency also requires that industries locate in those regions that offer them the highest possible productivity. Structural transformation allows regions to accommodate all or part of their differences in industry productivities without productivity deepening. If all regions have enough industries with the highest productivity, factor prices are equalized across regions. If some regions do not have enough industries with the highest productivity, they are forced to produce in low productivity industries and must lower their factor prices to be able to compete internationally. Subsection 2.5 shows how this affects the properties of the growth process.

In the presence of these two types of inefficiency, diminishing returns retain a regional scope even after globalization. Regional differences in factor prices still reflect regional differences in factor abundance and industry productivities, although the mapping between these variables is much more subtle than in the world of autarky. However, even in the presence of these inefficiencies regional differences in factor prices cannot reflect regional differences in market size. For market size effects to retain a regional scope after globalization we need to introduce impediments to goods trade. And this task is left for section 3.

2.4 Limits to structural transformation (I): factor proportions

It follows from Definition (36) that factor prices are equalized if and only if it is possible to achieve full employment of human and physical capital in all regions producing only with the highest productivity (requirement R1), with the factor proportions used in the integrated economy (requirement R2), and without incurring

a fixed cost more than once (requirement R3). Moving away from the factor-price-equalization world means that we must consider the violation of one or more of these requirements. Since the market for each input is "small", I assume that regions are large enough to ensure that requirement R3 is always satisfied.[38] Therefore, in the remainder of this section I will focus on violations of requirements R1 and R2. In this subsection, we study the effects on the growth process of violations to requirement R2, keeping the assumption that requirement R1 is not binding. This assumption will be removed in sub-section 2.5.

To formalize the notion that requirement R1 is not binding, define $I^*_{c,t}$ as the set of industries in which region c has the highest possible productivity: $I^*_{c,t} \equiv \left\{ \mathrm{i} \in I \;\middle|\; \mathrm{c} \in \underset{\mathrm{c'} \in C}{\arg\max} \left\{ Z_{\mathrm{c'},\mathrm{it}} \right\} \right\}$ for all $c \in C$. To ensure that requirement R1 is not binding in the models of this section, for each of them I first construct the set of "unrestricted" world equilibria by assuming that $I^*_{c,t} = I$ for all $c \in C$. As mentioned, all these equilibria share the same prices and world aggregates, but might exhibit different geographical patterns of production. In these "unrestricted" world equilibria, some industries might not operate in all regions. Naturally, prices and world aggregates would not be affected if regions did not have the best possible technologies in some or all of the industries in which they do not produce. Therefore, we can trivially relax the assumption that $I^*_{c,t}$ contains all industries, and instead assume only that there exists an "unrestricted" equilibrium such that, for all $c \in C$, the industries not included in $I^*_{c,t}$ do not operate in the region. This defines the extent to which regional differences in industry productivities are allowed in this section. It follows that requirement R1 is never binding and comparative advantage is determined solely by regional differences in factor proportions.

[38] I shall explore the effects of violations to requirement R3 in sections 3.2 and 3.3.

In the worlds we consider in this sub-section it is not possible in general to employ all factors in all regions using the techniques of the integrated economy. Even if they concentrate all of their production in industries that use human capital intensively, regions with abundant human capital might lack enough physical capital to produce with the factor proportions that these industries would use in the integrated economy. These regions are therefore forced to use a higher proportion of human capital in their industries and this requires them to have a lower wage-rental ratio than in the integrated economy. Naturally, the exact opposite occurs in regions with abundant physical capital. This situation can be aptly described as a geographical mismatch between different factor endowments.

To study the causes and effects of this mismatch, I present two examples that help build intuitions that apply more generally. The first example is the two-industry case that is so popular in trade theory:

Example 2.4.1: Consider a world economy with H- and K-industries, $I^H \cup I^K = I$ and $I^H \cap I^K = \varnothing$. Assume $\alpha_i=\alpha_H$, $\sigma_i=\sigma_H$ and $\max_{c\in C}\{Z_{c,it}\} = Z_{Ht}$ if $i\in I^H$, $\alpha_i=\alpha_K$, $\sigma_i=\sigma_K$ and $\max_{c\in C}\{Z_{c,it}\} = Z_{Kt}$ if $i\in I^K$, with $\alpha_H \leq \alpha_K$. (Note that $I^H\cdot\sigma_H + I^K\cdot\sigma_K=1$) For simplicity, assume also that $\varepsilon_i=\varepsilon$ and $\beta_i=\beta$ for all $i\in I$. The first step is to relate prices and world income to production:[39]

$$(47) \quad P_{it} = \sigma_i \cdot \frac{\prod_{i'\in I}\left(\sum_{c\in C} Q_{c,i't}\right)^{\sigma_i}}{\sum_{c\in C} Q_{c,it}} \qquad \text{for all } i\in I$$

$$(48) \quad Y_t = \prod_{i\in I}\left(\sum_{c\in C} Q_{c,it}\right)^{\sigma_i}$$

[39] These Equations follow from Equations (3) and (4).

Equation (47) can be thought of as the "demand" side of the model, since it shows how prices depend negatively on quantities, while Equation (48) simply describes world income. The "supply" side of the model is given by the following set of Equations:[40]

$$(49)\quad \frac{(1-\alpha_k)\cdot P_{Kt}}{w_{c,t}}\cdot\sum_{i\in I^K} Q_{c,it}+\frac{(1-\alpha_H)\cdot P_{Ht}}{w_{c,t}}\cdot\sum_{i\in I^H} Q_{c,it}=H_{c,t} \qquad \text{for all } c\in C$$

$$(50)\quad \frac{\alpha_k\cdot P_{Kt}}{r_{c,t}}\cdot\sum_{i\in I^K} Q_{c,it}+\frac{\alpha_H\cdot P_{Ht}}{r_{c,t}}\cdot\sum_{i\in I^H} Q_{c,it}=K_{c,t} \qquad \text{for all } c\in C$$

$$(51)\quad \left(\frac{w_{c,t}}{1-\alpha_H}\right)^{1-\alpha_H}\cdot\left(\frac{r_{c,t}}{\alpha_H}\right)^{\alpha_H}\geq\frac{\varepsilon-1}{\varepsilon}\cdot Z_{Ht}\cdot p_{Ht}=\left(A_{Ht}\cdot P_{Ht}\right)^{\frac{1}{\mu}}\cdot\left(\sigma_H\cdot Y_t\right)^{\frac{\mu-1}{\mu}}=\phi_H\cdot f_{Ht} \quad \text{for all } c\in C$$

$$(52)\quad \left(\frac{w_{c,t}}{1-\alpha_K}\right)^{1-\alpha_K}\cdot\left(\frac{r_{c,t}}{\alpha_K}\right)^{\alpha_K}\geq\frac{\varepsilon-1}{\varepsilon}\cdot Z_{Kt}\cdot p_{Kt}=\left(A_{Kt}\cdot P_{Kt}\right)^{\frac{1}{\mu}}\cdot\left(\sigma_K\cdot Y_t\right)^{\frac{\mu-1}{\mu}}=\phi_K\cdot f_{Kt} \qquad \text{for all } c\in C$$

where $\phi_i=(1-\alpha_i)^{\alpha_i-1}\cdot\alpha_i^{-\alpha_i}$ for all $i\in I$; and f_{Ht} and f_{Kt} are measures of the lowest factor costs in the world for the H- and K-industries since in equilibrium $f_{it}=\min_{c\in C}\left\{w_{c,t}^{1-\alpha_i}\cdot r_{c,t}^{\alpha_i}\right\}$ for all $i\in I$. Equations (49)-(50) are factor market clearing conditions, while Equations (51)-(52) are just a transformation of the pricing equations of each industry (for both final goods and intermediate inputs). Naturally, these pricing equations hold with strict equality if there is positive production in the corresponding industry. Equations (49)-(52) determine the production of each type of industry and the factor prices of region c, as a function of world prices and income.[41]

[40] Equations (49)-(50) follow from Equations (45) and (46), while Equations (51)-(52) follow from Equations (7) and (9) after using Equation (17) to eliminate the number of input varieties.

[41] If one is willing to take goods prices and factor endowments parametrically and further assume that the pricing equations hold with strict equality, it is possible to derive two popular results of trade theory from Equations (49)-(52). The Stolper-Samuelson effect says that an increase in the relative price of an industry leads to a more than proportional increase in the price of the factor that is used intensively in this industry and a decline in the price of the other factor. The Rybcynski effect says that an increase in a factor endowment leads to a more than proportional increase in the production of the industry that uses this factor intensively and a decline in the production of the other industry.

Equations (47)-(52) determine prices and quantities as a function of the distribution of factor endowments. Together with the regional analogues to Equation (1), the initial condition and the dynamics of the exogenous state variables, these Equations provide a complete characterization of the world equilibrium. Next, I describe some its most salient features.

Regions with extreme factor proportions have specialized production structures, while regions with intermediate factor proportions have diversified production structures. Let C_{Kt} (C_{Ht}) be the set of regions where there is production only in K-industries (H-industries), and let C_{Mt} be the set of regions where there is production in both types of industries. In fact, it follows from Equations (49)-(52) that these sets of regions are defined as follows:

$$(53) \quad C_{Kt} = \left\{ c \in C \;\middle|\; \frac{H_{c,t}}{K_{c,t}} \leq \frac{1-\alpha_K}{\alpha_K} \cdot \left(\frac{f_{Kt}}{f_{Ht}} \right)^{\frac{1}{\alpha_K - \alpha_H}} \right\}$$

$$(54) \quad C_{Ht} = \left\{ c \in C \;\middle|\; \frac{H_{c,t}}{K_{c,t}} \geq \frac{1-\alpha_H}{\alpha_H} \cdot \left(\frac{f_{Kt}}{f_{Ht}} \right)^{\frac{1}{\alpha_K - \alpha_H}} \right\}$$

$$(55) \quad C_{Mt} = \left\{ c \in C \;\middle|\; \frac{1-\alpha_K}{\alpha_K} \cdot \left(\frac{f_{Kt}}{f_{Ht}} \right)^{\frac{1}{\alpha_K - \alpha_H}} < \frac{H_{c,t}}{K_{c,t}} < \frac{1-\alpha_H}{\alpha_H} \cdot \left(\frac{f_{Kt}}{f_{Ht}} \right)^{\frac{1}{\alpha_K - \alpha_H}} \right\}$$

It follows from Equations (51)-(52) that factor prices are the same for all $c \in C_M$. If the dispersion in regional factor proportions is not too large, and the dispersion in factor intensities is not too low, C_{Kt}=C_{Ht}=$\varnothing$ and there is factor price equalization. Otherwise, this world economy exhibits a limited version of the factor price equalization result since factor prices are still equalized for all $c \in C_{Mt}$. It is common in

trade theory to refer to a group of regions that share the same factor prices as a "cone of diversification". In fact, we can write the wage and the rental as a function of f_{Ht} and f_{Kt} as follows:

$$(56) \quad w_{c,t} = \begin{cases} (1-\alpha_K)\cdot\phi_K \cdot f_{Kt} \cdot \left(\frac{H_{c,t}}{K_{c,t}}\right)^{-\alpha_K} & \text{if } c \in C_{Kt} \\ f_{Kt}^{\frac{-\alpha_H}{\alpha_K-\alpha_H}} \cdot f_{Ht}^{\frac{\alpha_K}{\alpha_K-\alpha_H}} & \text{if } c \in C_{Mt} \\ (1-\alpha_H)\cdot\phi_H \cdot f_{Ht} \cdot \left(\frac{H_{c,t}}{K_{c,t}}\right)^{-\alpha_H} & \text{if } c \in C_{Ht} \end{cases}$$

$$(57) \quad r_{c,t} = \begin{cases} \alpha_K \cdot \phi_K \cdot f_{Kt} \cdot \left(\frac{H_{c,t}}{K_{c,t}}\right)^{1-\alpha_K} & \text{if } c \in C_{Kt} \\ f_{Kt}^{\frac{1-\alpha_H}{\alpha_K-\alpha_H}} \cdot f_{Ht}^{\frac{\alpha_K-1}{\alpha_K-\alpha_H}} & \text{if } c \in C_{Mt} \\ \alpha_H \cdot \phi_H \cdot f_{Ht} \cdot \left(\frac{H_{c,t}}{K_{c,t}}\right)^{1-\alpha_H} & \text{if } c \in C_{Ht} \end{cases}$$

The wage is continuous and weakly declining on the human to physical capital ratio, while the rental is also continuous but increasing on this same ratio. The most noteworthy feature of these relationships is that they exhibit a "flat" for the set of human to physical capital ratios that define the cone of diversification. The top panel of Figure 10 shows how the wage-rental varies with a region's ratio of human to physical capital. Regional differences in this ratio reflect factor abundance in the usual way. In regions with a high (low) ratio of human to physical capital the price of human capital is low (high) relative to physical capital. Factor prices do not reflect however regional differences in industry productivities and/or market size.

It is now straightforward to compute the world income distribution as a function of f_{Ht} and f_{Kt}:

$$(58)\quad Y_{c,t}=\begin{cases}\phi_K\cdot f_{Kt}\cdot H_{c,t}^{1-\alpha_K}\cdot K_{c,t}^{\alpha_K} & \text{if } c\in C_{Kt}\\ f_{Kt}^{\frac{-\alpha_H}{\alpha_K-\alpha_H}}\cdot f_{Ht}^{\frac{\alpha_K}{\alpha_K-\alpha_H}}\cdot H_{c,t}+f_{Kt}^{\frac{1-\alpha_H}{\alpha_K-\alpha_H}}\cdot f_{Ht}^{\frac{\alpha_K-1}{\alpha_K-\alpha_H}}\cdot K_{c,t} & \text{if } c\in C_{Mt}\\ \phi_H\cdot f_{Ht}\cdot H_{c,t}^{1-\alpha_H}\cdot K_{c,t}^{\alpha_H} & \text{if } c\in C_{Ht}\end{cases}$$

We can use Equation (58) to re-evaluate earlier results about the relative contribution of factor endowments and industry productivities to income differences across regions. The first result is that the elasticity of substitution between human and physical capital is one outside the cone of diversification, but infinity within the cone. This elasticity reflects the relative importance of structural transformation and factor deepening as means to absorb regional differences in factor proportions. The second result is that regional differences in industry productivities continue not playing a role in determining regional income differences. This, of course, is not surprising given the assumption we have made about requirement R1 not being binding. The third and final result is that relative incomes are homogenous of degree one on factor endowments. This not surprising either since it simply confirms the absence of market size effects at the regional level.

We can also write the dynamics of the capital stock as a function of f_{Ht} and f_{Kt}:

$$(59)\quad K_{c,t+1}=\begin{cases}S_{c,t}\cdot(1-\alpha_K)\cdot\phi_K\cdot f_{Kt}\cdot H_{c,t}^{1-\alpha_K}\cdot K_{c,t}^{\alpha_K} & \text{if } c\in C_{Kt}\\ S_{c,t}\cdot f_{Kt}^{\frac{-\alpha_H}{\alpha_K-\alpha_H}}\cdot f_{Ht}^{\frac{\alpha_K}{\alpha_K-\alpha_H}}\cdot H_{c,t} & \text{if } c\in C_{Mt}\\ S_{c,t}\cdot(1-\alpha_H)\cdot\phi_H\cdot f_{Ht}\cdot H_{c,t}^{1-\alpha_H}\cdot K_{c,t}^{\alpha_H} & \text{if } c\in C_{Ht}\end{cases}$$

The specific dynamics of this example are hard to determine, since f_{Ht} and f_{Kt} change from generation to generation. It is easy to construct examples in which the world economy moves towards factor-price equalization; examples in which the world economy moves away from factor-price equalization; or examples in which the world economy alternates between periods in which factor prices are equalized and periods in which they are not. These dynamics depend on all the parameters the

model (including initial condition) and the evolution of the exogenous state variables, i.e. savings, human capital and industry productivities. Regardless of the specific dynamics, the world income distribution is stable if the joint distribution of these variables is stationary. Economic growth is positively transmitted across regions through changes in goods prices. This stabilizing role of trade is further reinforced by the fact that regions outside the cones cannot absorb capital accumulation through structural transformation and, consequently, experience diminishing returns in production.

Identifying cones of diversification is important because regional differences in factor proportions lead to structural transformation inside them, but to factor deepening outside them. In example 2.4.1, there is one of such cones and contains regions with intermediate factor proportions. Regions with extreme factor proportions do not belong to any cone. This need not be always the case, as the next example shows.

Example 2.4.2: Consider a world economy with H-, M- and K-industries, $I^H \cup I^M \cup I^K = I$, $I^H \cap I^M = \varnothing$, $I^H \cap I^K = \varnothing$ and $I^M \cap I^K = \varnothing$. Assume α_i=0 and $\max_{c \in C}\{Z_{c,it}\} = Z_{Ht}$ if $i \in I^H$; α_i=α_M and $\max_{c \in C}\{Z_{c,it}\} = Z_{Mt}$ if $i \in I^M$; and α_i=1 and $\max_{c \in C}\{Z_{c,it}\} = Z_{Kt}$ if $i \in I^K$. For simplicity, assume also that ε_i=ε and β_i=β for all $i \in I$. The "demand" side of this model is still described by Equations (47)-(48), but the "supply side is now given by:

(60) $$\frac{(1-\alpha_M)\cdot P_{Mt}}{w_{c,t}} \cdot \sum_{i \in I^M} Q_{c,it} + \frac{P_{Ht}}{w_{c,t}} \cdot \sum_{i \in I^H} Q_{c,it} = H_{c,t} \qquad \text{for all } c \in C$$

(61) $$\frac{P_{Kt}}{r_{c,t}} \cdot \sum_{i \in I^K} Q_{c,it} + \frac{\alpha_M \cdot P_{Mt}}{r_{c,t}} \cdot \sum_{i \in I^M} Q_{c,it} = K_{c,t} \qquad \text{for all } c \in C$$

(62) $$w_{c,t} \geq \frac{\varepsilon - 1}{\varepsilon} \cdot Z_{Ht} \cdot p_{Ht} = \left(A_{Ht} \cdot P_{Ht}\right)^{\frac{1}{\mu}} \cdot \left(\sigma_H \cdot Y_t\right)^{\frac{\mu-1}{\mu}} = f_{Ht} \quad \text{for all } c \in C$$

(63) $$\left(\frac{w_{c,t}}{1-\alpha_M}\right)^{1-\alpha_M} \cdot \left(\frac{r_{c,t}}{\alpha_M}\right)^{\alpha_M} \geq \frac{\varepsilon - 1}{\varepsilon} \cdot Z_{Mt} \cdot p_{Mt} = \left(A_{Mt} \cdot P_{Mt}\right)^{\frac{1}{\mu}} \cdot \left(\sigma_M \cdot Y_t\right)^{\frac{\mu-1}{\mu}} = \phi_M \cdot f_{Mt} \quad \text{for all } c \in C$$

(64) $$r_{c,t} \geq \frac{\varepsilon - 1}{\varepsilon} \cdot Z_{Kt} \cdot p_{Kt} = \left(A_{Kt} \cdot P_{Kt}\right)^{\frac{1}{\mu}} \cdot \left(\sigma_K \cdot Y_t\right)^{\frac{\mu-1}{\mu}} = f_{Kt} \quad \text{for all } c \in C$$

Unlike the previous example, we find now that regions with extreme factor proportions have diversified production structures, while regions with intermediate factor proportions have specialized production structures. These sets of regions are now given by:[42]

(65) $$C_{Kt} = \left\{ c \in C \,\middle|\, \frac{H_{c,t}}{K_{c,t}} \leq \frac{1-\alpha_M}{\alpha_M} \cdot \left(\frac{f_{Kt}}{f_{Mt}}\right)^{\frac{1}{1-\alpha_M}} \right\}$$

(66) $$C_{Ht} = \left\{ c \in C \,\middle|\, \frac{H_{c,t}}{K_{c,t}} \geq \frac{1-\alpha_M}{\alpha_M} \cdot \left(\frac{f_{Mt}}{f_{Ht}}\right)^{\frac{1}{\alpha_M}} \right\}$$

(67) $$C_{Mt} = \left\{ c \in C \,\middle|\, \frac{1-\alpha_M}{\alpha_M} \cdot \left(\frac{f_{Kt}}{f_{Mt}}\right)^{\frac{1}{1-\alpha_M}} < \frac{H_{c,t}}{K_{c,t}} < \frac{1-\alpha_M}{\alpha_M} \cdot \left(\frac{f_{Mt}}{f_{Ht}}\right)^{\frac{1}{\alpha_M}} \right\}$$

Regions in C_{Kt} (C_{Ht}) produce in the M-industries and the K-industries (H-industries), while regions in C_{Mt} produce only in M-industries. Factor prices are determined as follows:

[42] Note that the sets C_{Kt} and C_{Ht} never intersect in world equilibrium. Assume the opposite, then it follows that equilibrium input prices must satisfy $f_{Mt} < (f_{Ht})^{1-\alpha_M} \cdot (f_{Kt})^{\alpha_M}$. But if this inequality held nobody would produce in M-industries and markets for the products of these industries would not clear.

$$(68)\qquad w_{c,t} = \begin{cases} f_{Mt}^{\frac{1}{1-\alpha_M}} \cdot f_{Ht}^{\frac{-\alpha_M}{1-\alpha_M}} & \text{if } c \in C_{Kt} \\ (1-\alpha_M)\cdot\phi_M \cdot f_{Mt} \cdot \left(\frac{H_{c,t}}{K_{c,t}}\right)^{-\alpha_M} & \text{if } c \in C_{Mt} \\ f_{Ht} & \text{if } c \in C_{Ht} \end{cases}$$

$$(69)\qquad r_{c,t} = \begin{cases} f_{Kt} & \text{if } c \in C_{Kt} \\ \alpha_M \cdot \phi_M \cdot f_{Mt} \cdot \left(\frac{H_{c,t}}{K_{c,t}}\right)^{1-\alpha_M} & \text{if } c \in C_{Mt} \\ f_{Mt}^{\frac{1}{\alpha_M}} \cdot f_{Ht}^{\frac{\alpha_M - 1}{\alpha_M}} & \text{if } c \in C_{Ht} \end{cases}$$

Once again, the wage is continuous and weakly declining on the human to physical capital ratio, while the rental is also continuous but increasing on this same ratio. But now these relationships exhibit at most two "flats", one for each set of human to physical capital ratios that defines a cone of diversification. Regional differences in factor prices reflect again factor abundance in the usual way. This world economy contains at most two cones of diversification.[43] Regions with extreme factor proportions belong to one of them, while regions with intermediate factor proportions do not. The middle panel of Figure 10 shows how the wage-rental varies with a region's ratio of human to physical capital.

It is straightforward to compute the analogues of Equation (58)-(59) for this example and check that the mapping from factor endowments to incomes and capital accumulation is also linear within the cones and takes the Cobb-Douglas form outside of them. The picture of the growth process that comes out of this example is therefore very similar to the on in Example 2.4.1.

[43] I say "at most" because it is also possible that $\underline{R}_{Mt} = \overline{R}_{Mt}$, in which the case there would be a single cone. Cuñat and Mafezzoli [2004a] analyze a similar model under the assumption that none of the regions of the world have specialized production structures, i.e. C_{Mt}=∅.

Examples 2.4.1 and 2.4.2 can be generalized by introducing further industries with different factor intensities. As we do so, the potential number of cones increases. But the overall picture remains the same. The world economy sorts itself out in a series of cones of diversification. The bottom panel of Figure 10 depicts a case with multiple cones of diversification.[44] Small regional differences in factor proportions lead to structural transformation within cones, but to factor deepening outside them. Large regional differences in factor proportions might span one or more cones and therefore lead to a mix of structural transformation and factor deepening. Therefore, the world of diversification cones can be seen as being somewhere in between the world of factor-price-equalization and the world of autarky.[45]

In the light of these results, we must slightly revise our earlier discussion of the effects of globalization on the source of income differences. As in the world of factor-price equalization, differences in domestic productivities cannot be a source of income differences and relative incomes are homogeneous of degree one with respect to factor endowments. But unlike the world of factor-price-equalization, the elasticity of substitution between domestic factors is no longer infinity but instead lies somewhere between one (outside cones) and infinity (within cones). As mentioned, this elasticity measures the relative importance of structural transformation and factor deepening as a means to accommodate regional differences in factor proportions. And this relative importance in turn depends on various factors, most notably how dispersed are factor intensities across industries. Two extreme examples make this point forcefully. If the dispersion in industry factor intensities is extreme, i.e. $\alpha_i \in \{0,1\}$ for all $i \in I$, then regional differences in factor proportions always lead to structural transformation and the world income distribution is given by

[44] Dornbusch, Fischer and Samuelson [1980] develop a similar model with a continuum of goods that vary in their factor intensity, although they do not specifically study the formation of cones.
[45] In pure Heckscher-Ohlin models, Deardorff [2001] and Cuñat and Maffezzoli [2004a] generate "club convergence". Stiglitz [1970] and Devereux and Shi [1991] are examples where cones of diversification establish due to inherently different time-preferences and incomes diverge. Oniki and Uzawa [1965] analyze conditions for diversification cones in two-sector model.

Equations (39)-(40).[46] If the dispersion in industry factor intensities is instead negligible, i.e. $\alpha_i=\alpha$ for all $i \in I$, then regional differences in factor proportions always lead to factor deepening and the world income distribution is given by:[47]

(70) $$Y_{c,t} = A_t \cdot H_{c,t}^{1-\alpha} \cdot K_{c,t}^{\alpha}$$

As in the world of autarky, the elasticity of substitution across factors is one (see Equation (33)). But unlike the world of autarky, regional differences in industry productivities and market size play no role in explaining regional income differences.

We do not need to revise however our earlier discussion of how globalization affects the dynamic responses to region-specific shocks. In this respect, the world with diversification cones offers the same insights as the world of factor-price-equalization. Region-specific shocks to savings and human capital have positive effects that spill over to other regions, while shocks to industry productivities only have effects if they push outwards the world productivity frontier. Economic growth is positively transmitted across regions through changes in goods prices and this keeps the world income distribution stable. In fact, this force towards stability is further reinforced in regions that are outside a cone by the existence of diminishing returns in production.

We conclude therefore that violations to requirement R2 do not alter much the picture came out of the factor-price-equalization world. Surely the geographical mismatch between different factor endowments implied by these violations might generate large inefficiencies that, in turn, might lead to sizeable regional differences in factor prices. Therefore, there might be important quantitative differences between a world with many diversification cones and the world of factor-price-equalization. But the qualitative properties of the growth process of these two worlds remain relatively close to each other, and far away from those of the world of autarky.

[46] This is the model used by Ventura [1997].
[47] One of many ways to find this result is as the appropriate limiting case of Examples 2.4.1 or 2.4.2.

2.5 Limits to structural transformation (II): industry productivities

Consider next worlds where requirement R1 is either binding or fails. Regions with few high-productivity industries might find that even if they concentrate all of their production in those industries, they cannot employ all of their factors and produce the same quantities as the integrated economy. These regions are therefore forced to exceed the production of the integrated economy in those industries and/or move into low-productivity industries. Whatever the case, this requires these regions to offer low factor prices to employ all of their factors. This situation can be aptly described as a geographical mismatch between industry productivities and factor endowments.

To make further progress, it is necessary to be more explicit about why and how industry productivities differ across regions. The first example considers the case in which regional differences in productivities take the popular factor-augmenting form:

Example 2.5.1: Consider a world where $Z_{c,it} = \pi_{c,Ht}^{1-\alpha_i} \cdot \pi_{c,Kt}^{\alpha_i}$ for all $i \in I$ and all $c \in C$; with $\sum_{c \in C} \pi_{c,Ht} \cdot \frac{H_{c,t}}{H_t} = 1$ and $\sum_{c \in C} \pi_{c,Kt} \cdot \frac{K_{c,t}}{K_t} = 1$. As usual, $\pi_{c,Ht}$ and $\pi_{c,Kt}$ are interpreted as labor- and capital-augmenting productivity differences, The world productivity frontier is given by $Z_{it} = \max_c \{\pi_{c,Ht}^{1-\alpha_i} \cdot \pi_{c,Kt}^{\alpha_i}\}$. In the integrated economy, industries would be located exclusively in the regions that are in this frontier. The set FPE_t is "small" and, except for a few very special or knife-edge cases, factor-price equalization is not possible and requirement R1 fails. [48]

[48] Take, for instance, the case of two regions and two industries. If one region has the highest productivity in both industries the only factor distribution that leads to factor-price equalization is the one

To understand the logic of this world, it is useful to follow the usual procedure of re-normalizing the model in terms of "efficiency" or "productivity-equivalent" factor units. That is, we can pretend that regional factor endowments are given by $\hat{H}_{c,t} = \pi_{c,Ht} \cdot H_{c,t}$ and $\hat{K}_{c,t} = \pi_{c,Kt} \cdot K_{c,t}$ for all $c \in C$; and that industry productivities are identical across regions, i.e. $\hat{Z}_{c,it} = 1$ for all $i \in I$ and all $c \in C$. Then, productivity-adjusted factor prices are given by $\hat{w}_{c,t} = \frac{w_{c,t}}{\pi_{c,Ht}}$ and $\hat{r}_{c,t} = \frac{r_{c,t}}{\pi_{c,Kt}}$. The key observation is that the re-normalized model is formally equivalent to the model of the previous section.[49] Therefore, all the results we obtained in the previous sections regarding the cross-section of factor prices also apply here to productivity-adjusted factor prices, i.e. $\hat{w}_{c,t}$ and $\hat{r}_{c,t}$; but not to factor prices as usually measured, i.e. $w_{c,t}$ and $r_{c,t}$.[50]

As the worlds of the previous section, this world economy sorts itself out in a series of cones of diversification. All regions within a cone have the same productivity-adjusted factor prices, although possibly different factor prices as usually measured. Regional differences in productivity-adjusted factor proportions lead to structural transformation within cones, and to factor deepening across them. When all regions are located within a single cone, we have the conditional factor-price-equalization result emphasized by Trefler [1993]. That is, regional differences in factor prices reflect only differences in factor-augmenting productivities and are not related to differences in productivity-adjusted factor proportions.

---

in which all factors are located in this region. If instead each region has the highest productivity in a different industry, the only factor distribution that leads to factor-price equalization is the one in which each region receives the exact quantity of factors that its high-productivity industry uses in the integrated economy.

[49] The re-normalized model is a bit less general than the model of the previous section since it does not display regional differences in industry productivities. We could (trivially) generalize this example to allow for regional differences in industry productivities, but keeping the assumption that requirement R1 is not binding in the re-normalized model.

[50] For instance, Equations (56)-(57) describe the productivity-adjusted factor if we further assume that the world economy contains two types of industries as in Example 2.4.1. Similarly, Equations (68)-(69) describe productivity-adjusted factor prices if we instead assume that the world economy contains three types of industries as in Example 2.4.2.

Although the presence of factor-augmenting productivity differences does not alter much the formal or mathematical structure of the model, it has important implications for the question of why some regions are richer than others. Unlike the worlds of section 2.4, we now have that productivity differences become a source of income differences across countries. For instance, if all regions belong to a single cone of diversification we have the following counterpart to Equation (40):

$$(71) \qquad \frac{Y_{c,t}}{Y_t} = (1-\alpha)\cdot\frac{\pi_{c,Ht}\cdot H_{c,t}}{H_t} + \alpha\cdot\frac{\pi_{c,Kt}\cdot K_{c,t}}{K_t} \qquad \text{for all } c\in C$$

Alternatively, if all the industries in the world have the same factor intensity we have the following counterpart of Equation (70):

$$(72) \qquad Y_{c,t} = A_{c,t}\cdot H_{c,t}^{1-\alpha}\cdot K_{c,t}^{\alpha}$$

where $A_{c,t} = \hat{A}_t\cdot\pi_{c,Ht}^{1-\alpha}\cdot\pi_{c,Kt}^{\alpha}$.[51] The inability of the world economy to match best technologies with appropriate factors moves us a step closer to the world of autarky, since regional productivities now affect regional incomes. Moreover, since now the world operates below its productivity frontier shocks to regional factor productivities have effects even if they do not push this frontier. Note however that, as in the worlds of section 2.4, the elasticity of substitution between domestic factors still lies somewhere between one (outside cones) and infinity (within cones); and relative incomes are homogeneous of degree one with factor endowments.

The rest of the picture of the growth process that comes out of this world remains close to the world of factor-price-equalization. Region-specific shocks to savings and human capital have positive effects that spill over to other regions. Economic growth is positively transmitted across regions through goods prices and

[51] This model therefore provides an alternative theoretical foundation for the work of Mankiw, Romer and Weil [1992], Hall and Jones [1999] and Klenow and Rodríguez-Clare [1997].

this keeps the world income distribution stable. If the conditional version of the factor-price-equalization theorem does not hold, regions outside the cones experience diminishing returns and this reinforces the effects of changes in product prices on the stability of the world income distribution.

Assuming that regional productivity differences take the factor-augmenting form discussed in example 2.5.1 is popular because it yields tractable models. But the factor-augmenting view of productivity differences hides some interesting effects of trade on the world income distribution and its stability. One reason is that, in the world of factor-augmenting productivity differences, comparative advantage is still determined solely by regional differences in factor proportions, albeit productivity-adjusted ones. The next example provides a dramatic illustration of how regional differences in industry productivities could determine comparative advantage, and how this brings about a new effect of trade on the world income distribution:

Example 2.5.2: Consider a world with many industries and regions. Assume that $Z_{c,it} = 1$ if $i \in I^*_{c,t}$, and $Z_{c,it} = 0$ if $i \notin I^*_{c,t}$; where $I^*_{c,t}$ for all $c \in C$ constitutes a partition of I: $\bigcup_{c \in C} I^*_{c,t} = I$ and $I^*_{c,t} \cap I^*_{c',t} = \varnothing$ for all $c \in C$ and $c' \in C$. Assume also that $I^*_{c,t} \neq \varnothing$ for all $c \in C$. That is, each region knows how to produce a disjoint subset of goods. Since only one region knows how to produce each good, the corresponding industry is located in that region. That is, $I_{c,t} = I^*_{c,t}$ for all $c \in C$, regardless of the factor distribution. In this world, comparative advantage is driven solely by regional differences in industry productivities, and differences in factor proportions play no role. In this example, requirement R1 does not fail but it is binding, except for a few very special and knife-edge cases.

A bit of straightforward algebra shows that production and factor allocations are given as follows:

(73) $$Y_{c,t} = \sum_{i \in I^*_{c,t}} \phi_i \cdot f_{it} \cdot H_{c,it}^{1-\alpha_i} \cdot K_{c,it}^{\alpha_i} \quad \text{for all } c \in C$$

(74) $$H_{c,it} = \frac{\sigma_i}{\sum_{i' \in I^*_{c,t}} \sigma_{i'}} \cdot \frac{1-\alpha_i}{\sum_{i' \in I^*_{c,t}} \sigma_{i'} \cdot (1-\alpha_{i'})} \cdot H_{c,t} \text{ if } i \in I^*_{c,t}; \text{ and } H_{c,it} = 0 \text{ if } i \notin I^*_{c,t}$$

(75) $$K_{c,it} = \frac{\sigma_i}{\sum_{i' \in I^*_{c,t}} \sigma_{i'}} \cdot \frac{\alpha_i}{\sum_{i' \in I^*_{c,t}} \sigma_{i'} \cdot \alpha_{i'}} \cdot K_{c,t} \text{ if } i \in I^*_{c,t}; \text{ and } K_{c,it} = 0 \text{ if } i \notin I^*_{c,t}$$

where, as usual by now, $\phi_i = (1-\alpha_i)^{\alpha_i - 1} \cdot \alpha_i^{-\alpha_i}$ and $f_{it} = \min_{c \in C} \left\{ w_{c,t}^{1-\alpha_i} \cdot r_{c,t}^{\alpha_i} \right\}$ for all $i \in I$.

Equation (73) describes the world income distribution as a function of factor allocations and goods prices, while Equations (74)-(75) provide the equilibrium factor allocations as a function of aggregate factor endowments. By substituting Equations (74)-(75) into Equation (73), we obtain the world income distribution as a function of factor endowments and input prices.[52] It is immediate to show that the elasticity of substitution between human and physical capital is between one and infinity; that regional differences in industry productivities affect regional differences in income; and that the world income distribution is homogeneous of degree one with respect to factor endowments.

These results are obtained from a relationship between incomes, factor endowments and industry productivities that holds constant input prices. Once we substitute input prices into this relationship, we find that the world income distribution is given by:

(76) $$\frac{Y_{c,t}}{Y_t} = \sum_{i \in I^*_{c,t}} \sigma_i \qquad \text{for all } c \in C$$

[52] This relationship is formally analogous, for instance, to Equation (58) in Example 2.4.1.

Equation (76) states that the share of world income of each region equals that share of world spending on the industries located in the region, and it does not depend on domestic factor endowments. What is going on? Assume a region has a ratio of human to physical capital λ times higher than average. Since the region is producing a fixed set of goods, it is forced to operate with a ratio that is λ times higher than average, and this requires a wage-rental ratio that is λ^{-1} higher than average. Therefore the elasticity of substitution between human and physical capital in production is one. What is different here is that relative incomes are now homogeneous of degree zero with respect to factor endowments. Assume a region's human and physical capitals are both λ times average. Since production is homogenous of degree one with factor endowments, its production of all industries is λ times average. But since the country faces a demand for its products with price-elasticity equal to one, the prices of its products are λ^{-1} times average. As a result, the income of the region is just average, despite its factor endowments being λ times average.

So what should we conclude about the degree of homogeneity of relative incomes with respect to factor endowments? As Equations (73)-(75) and (76) show, in empirical applications it will depend on whether we are holding goods prices constant or not. If we are holding these prices constant, then relative incomes are homogeneous of degree one in factor endowments. If we are not holding goods prices constant, then the degree of homogeneity of relative incomes with respect of factor endowments lies between zero and one. In this example, this degree of homogeneity is zero because regional differences in factor endowments are absorbed by regional variation in the quantities produced of each input. In Examples 2.4.1, 2.4.2 and 2.5.1, this degree of homogeneity was one because regional differences in factor endowments were absorbed by regional variation in the number of input varieties produced. The next example, inspired by Dornbusch, Fischer and Samuelson [1977], neatly clarifies this point by showing an intermediate world where both margins are at work.

Example 2.5.3: Consider a world with two regions, C={N,S}; and a continuum of industries, I=[0,1]. Assume all industries have the same factor intensity, $\alpha_i=\alpha$ for all $i \in I$. For simplicity, let also $\varepsilon_i=\varepsilon$ and $\beta_i=\beta$ for all $i \in I$. It follows immediately that: [53]

$$(77) \quad Y_{c,t} = \phi \cdot f_{c,t} \cdot H_{c,t}^{1-\alpha} \cdot K_{c,t}^{\alpha} \qquad \text{for all } c \in C$$

where $\phi = (1-\alpha)^{\alpha-1} \cdot \alpha^{-\alpha}$; and $f_{c,t}$ is a measure of factor costs of region c, i.e. $f_{c,t} = w_{c,t}^{1-\alpha} \cdot r_{c,t}^{\alpha}$ for all $c \in C$. To characterize the world income distribution in this world, we need to determine factor costs. Equation (77) is akin to Equation (58) or Equations (73)-(75) in the sense that it shows the world income distribution as a function of factor endowments and input prices. Not surprisingly, these relative incomes are homogeneous of degree one with respect to factor endowments. The next step is to determine input prices and substitute them into Equation (77).

Define $T_i \equiv \frac{Z_{N,it}}{Z_{S,it}}$ for all $i \in I$ as the industry productivity of North relative to South. Then, assign indices or order goods so that T_i is non-increasing in i. Note that T_i might be neither continuous nor invertible.[54] It follows from this ordering that $I_{N,t} \equiv \left\{ i \in I \;\middle|\; \frac{f_{N,t}}{f_{S,t}} \le T_i \right\}$ and $I_{S,t} \equiv \left\{ i \in I \;\middle|\; \frac{f_{N,t}}{f_{S,t}} \ge T_i \right\}$. That is, North (or N) specializes on low-index industries while South (or S) specializes in high-index industries. The cutoff industry, i*, is determined as follows:[55]

[53] This follows directly from the observation that the share of human and physical capital in income are 1-α and α, respectively.
[54] This ranking can vary over time, but this does not play any role here. Without loss of generality, the reader can focus on the case in which the ranking is time-invariant.
[55] If T_i is not invertible in the region of interest, this condition determines a set of candidate values for i*.

(78) $$\frac{f_{N,t}}{f_{S,t}} = T_{i^*}$$

Let X_i be world share of spending on all industries with indices equal or lower than i, that is, $X_i \equiv \int_0^i \sigma_j \cdot dj$. Note that X_i is non-decreasing in i, and takes values zero and one for i=0 and i=1. It follows from this definition that $Y_{N,t}=X_{i^*}\cdot(Y_{N,t}+Y_{S,t})$ and, using Equation (78), this can be rewritten as follows:

(79) $$\frac{f_{N,t}}{f_{S,t}} = \frac{X_{i^*}}{1-X_{i^*}} \cdot \left(\frac{H_{S,t}}{H_{N,t}}\right)^{1-\alpha} \cdot \left(\frac{K_{S,t}}{K_{N,t}}\right)^{\alpha}$$

Equations (78)-(79) jointly determine the pattern of production and trade (i*) and relative factor costs ($f_{N,t}/f_{S,t}$) as a function of spending patterns, industry productivities and factor endowments. Finally, we can use the numeraire rule in Equation (4) to find that:

(80) $$Y_t = \sum_{c\in\{N,S\}} \phi \cdot f_{c,t} \cdot H_{c,t}^{1-\alpha} \cdot K_{c,t}^{\alpha} = (\varepsilon - 1) \cdot \exp\left\{\int_0^{i^*} Z_{N,it} \cdot \sigma_i \cdot di + \int_{i^*}^{1} Z_{S,it} \cdot \sigma_i \cdot di\right\}$$

Having already found the pattern of production and trade (i*) and relative factor costs ($f_{N,t}/f_{S,t}$), Equation (80) can then be used to determine absolute factor costs.

This world is somewhat different form the ones we have seen so far in that we have only two regions. To think about the effects of factor endowments on relative incomes, I consider next a situation in which both regions have symmetric technologies and differ in that North's factor endowments are λ (>1) times larger than South's.[56] Figure 11 depicts this world. The AA and BB lines represent

[56] By symmetric technologies, I mean that if there exists an industry i such that $T_i=\tau$ then there also exists another industry i' such that $T_{i'}=1/\tau$ and $\alpha_i=\alpha_{i'}$, $\beta_i=\beta_{i'}$, $\varepsilon_i=\varepsilon_{i'}$ and $\sigma_i=\sigma_{i'}$.

Equations (78) and (79), respectively. The AA line is non-increasing because T_i is non-increasing in i, while the BB line is non-decreasing because X_i is non-decreasing in i. The existence of a unique crossing point follows since the BB line takes value zero at i=0 and slopes upward towards infinity at i=1.

The top panel of Figure 11 shows the case in which T_i is flat. This case corresponds to a world in which differences in industry productivities are minimal or irrelevant at the margin as in Examples 2.4.1, 2.4.2 and 2.5.1. This allows North to employ its larger factor endowments by producing a larger number of varieties than South. Factor costs are the same in both regions and, as a result, North's income is λ times South's. Relative incomes (after substituting in goods prices) are homogenous of degree one on factor endowments.

The middle panel of Figure 11 shows the opposite case in which T_i is vertical. This case corresponds to a world in which differences in industry productivities are extreme as in Examples 2.5.2. North is forced to employ its larger factor endowments by producing a higher quantity of each of its varieties. Factor costs in North are λ^{-1} times those of South and, as a result, North's income equals that of South. Relative incomes (after substituting in goods prices) are homogenous of degree zero on factor endowments.

The bottom panel shows the intermediate case in which T_i is neither flat nor vertical. Since the slope reflects how strong are differences in industry productivities, we are somewhere in between the two extreme examples considered up to now. North employs its larger factor endowments by producing a larger number of varieties and also a larger quantity of each of them. Factor costs in North are somewhere between λ^{-1} and one times those of South. The degree of homogeneity of relative incomes (after substituting in goods prices) on factor endowments is therefore somewhere between zero and one.

It is possible to generalize Example 2.5.3 in a variety of directions. For instance, one could allow for industry variation in factor intensities and many regions.[57] This is important in empirical applications, of course. But the central message remains. The effects of factor endowments on relative incomes depend on regional differences in industry productivities. If these differences are small, regions with larger factor endowments absorb them mostly through structural transformation: not changing much their production in existing industries and moving into new industries where the region's productivity relative to the rest of the world is similar to existing ones. If differences in industry productivities are large, regions with larger factor endowments absorb them by productivity deepening: substantially increasing their production in existing industries and/or moving into industries where the region's productivity relative to the rest of the world is substantially lower than in existing ones.

One can conclude from this discussion that differences in industry productivities create another force for diminishing returns to physical capital accumulation. As physical capital is accumulated, quantities produced increase and the terms of trade worsen. The result is a reduction in factor prices that lowers wages, savings and capital accumulation. This is a central aspect of the growth process in a world of interdependent regions generates a force towards the stability of the world income distribution.[58]

I argued at the end of section 1 that, if globalization leads to the integrated economy, there is a powerful prescription for economic development: open up and integrate into the world economy. This allows regions to benefit from higher productivity, improved factor allocation and increased market size. Not much has changed here. Naturally, if factor prices are equalized the effects are literally the same as in section 1 since the globalization leads to the integrated economy. If

[57] See Wilson [1980], Eaton and Kortum [1999, 2000], Matsuyama [2000] and Alvarez and Lucas [2004].
[58] See Acemoglu and Ventura [2002] on this point.

factor prices are not equalized, the world economy operates with a lower productivity and a worse factor allocation than the integrated economy. This also means that the size of the world economy will be smaller than that of the integrated economy. As a result, all the benefits from globalization are smaller in the worlds of sub-sections 2.4 and 2.5 than in the world of factor-price equalization. But it is still relatively straightforward to see that coupled with an appropriate transfer scheme globalization constitutes a Pareto improvement for the world economy. Moreover, since all regions gain from trade there exist Pareto-improving transfer schemes that do not require inter-regional transfers.[59] Therefore, the prescription for development remains the same: open up and integrate into the world economy.

We have traveled much already, and the global view of economic growth is starting to take shape. This view is more realistic and rich in details than the views that came out from either the world of autarky or the integrated economy. Despite this progress, we should not rest here yet. We have assumed so far that globalization eliminates all impediments to goods trade. This is obviously an unrealistic assumption. Is it also a crucial one?

3. Transport costs and market size

Despite the already large and growing importance of international trade, there are some important areas of economic activity where the degree of market integration is still relatively low. Surely the clearest case in point is the service sector.[60] As the textbook example of a haircut suggests, many services are

[59] How do we know that all regions have non-negative gains from globalization? Since regions have the choice of not trading, it is therefore possible to achieve the level of income and welfare of the world of autarky after globalization. Realizing these gains might require implementing an appropriate tax-subsidy scheme, though.

[60] In industrial economies, the service sector accounts for more than two thirds of production but only for about one fifth of exports and imports. Moreover, most trade in services is concentrated in activities related to transportation and travel even though these activities only constitute a small component of overall services production.

inherently more difficult to transport than agricultural and manufacturing products. Services also tend to be more vulnerable to various governmental barriers to trade, such as professional licensing requirements that discriminate against foreigners, domestic content requirements in public procurement, or poor protection of intellectual property rights.[61] In addition, there are important examples of weak market integration that go beyond the service sector. Trade in some agricultural and manufacturing products is also severely restricted as a result of protectionist practices in industrial countries.

The goal of this section is to study the effects on the growth process of partial segmentation in goods markets. The new model of globalization that I shall adopt here is as follows: at date t=0 the costs of transporting some (but not all) goods across regions suddenly fall from prohibitive to negligible. In particular, I partition the set of all industries into the sets of tradable and nontradable industries, i.e. T_t and N_t such that $T_t \cup N_t = I$ and $T_t \cap N_t = \varnothing$. The costs of transporting intermediate inputs and final goods fall from prohibitive to negligible at t=0 if $i \in T_t$. But even after t=0, the costs of transporting either the intermediate inputs, or the final goods, or both remain prohibitive if $i \in N_t$.[62] We keep assuming that the costs of transporting factors across regions remain prohibitive after t=0, and that international trade in assets is not possible. Naturally, the model analyzed in section 2 (and formally described in section 2.3) obtains as the special case of this model in which $T_t = I$ and $N_t = \varnothing$ for $t \geq 0$.

[61] There are also signs that this is changing rapidly. Advances in telecommunications technology, the appearance of e-commerce and the development of new and standardized software have all opened up the possibility of trading a wider range of services. Recent multilateral trade negotiations and the process of European integration have also led to the dismantling of various non-tariff barriers to service trade.

[62] The most popular alternative to this model is the "iceberg" cost model whereby all goods are subject to the same proportional transport cost. In particular, a quantity τ (>1) of a good must be shipped from source to ensure that one unit of it arrives to destination. The rest "melts" away in transit. See Matsuyama [2004] for yet another model of transport costs.

A central aspect of the analysis turns out to be whether transport costs apply only to final goods, to intermediate inputs, or to both. Section 3.1 presents the case in which transport costs apply only to final goods. This model neatly generalizes the results obtained in the previous section. Section 3.2 studies the case in which transport costs apply only to intermediate inputs. This gives rise to agglomeration effects that can have a large and somewhat unexpected impact on the world income distribution. Section 3.3 analyzes the case in which transport costs apply to both final goods and intermediate inputs. The interaction between the two types of frictions brings about a new perspective on the role of local markets.

3.1 Nontraded goods and the cost of living

Consider next a world where some final goods are not tradable, although the intermediate inputs required to produce them are always tradable. In particular, the costs of trading intermediate inputs are negligible for all $i \in I$; and the costs of transporting final goods are negligible if $i \in T_t$ but prohibitive if $i \in N_t$. Since prices of final goods can differ across regions, a novel feature of this model is that regions will have different price levels.

I must now revise the formal description of the model. While regional analogues of Equations (1) and (2) continue to apply, one must now recognize that final goods prices in nontradable industries might differ across regions. As a result, the price of consumption and investment will vary across them even if Equation (3) describing spending patterns still applies to all regions. Therefore, we must write the analogues of Equations (1) and (2) as follows:

(81) $$K_{c,t+1} = S_{c,t} \cdot \frac{w_{c,t}}{P_{c,t}} \cdot H_{c,t}$$

(82) $$C_{c,t} = (1 - S_{c,t}) \cdot \frac{w_{c,t}}{P_{c,t}} \cdot H_{c,t} + \frac{r_{c,t}}{P_{c,t}} \cdot K_{c,t}$$

where $P_{c,t}$ is the price level (or cost of living) of region c, i.e. $P_{c,t} = \prod_{i \in I} \left(\frac{P_{c,it}}{\sigma_i} \right)^{\sigma_i}$ for all $c \in C$. A natural choice of numeraire now is the ideal price index for tradable industries:

(83) $$1 = \prod_{i \in T_t} \left(\frac{P_{it}}{\sigma_i} \right)^{\sigma_i}$$

Equation (83) replaces Equation (4). The latter obtains as the special case of the former in which all goods are tradable, i.e. $T_t = I$ and $N_t = \varnothing$. An implication of this choice of numeraire is that the price level of region c is equal to the ideal price index of its nontradable industries:

(84) $$P_{c,t} = \prod_{i \in N_t} \left(\frac{P_{c,it}}{\sigma_i} \right)^{\sigma_i} \qquad \text{for all } c \in C$$

Since now price levels differ across regions it is necessary to distinguish between two concepts of income and factor prices: (1) market-based incomes and factor prices, i.e. $Y_{c,t}$, $w_{c,t}$ and $r_{c,t}$; and (2) real or PPP-adjusted incomes and factor prices, i.e. $Y_{c,t}/P_{c,t}$, $w_{c,t}/P_{c,t}$ and $r_{c,t}/P_{c,t}$. Whenever there is no risk of confusion, I shall refer to the former simply as income and factor prices, and to the latter as real income and real factor prices. As before, Equations (5)-(6) describing technology apply to all regions, with the corresponding factor prices and industry productivities.

After globalization, producers of intermediate inputs in all industries and producers of final goods in tradable industries face a global market and Equations

(7)-(10) describing pricing policies, input demands and the free-entry condition therefore apply only to those regions where the lowest-cost producers are located. But even after globalization, producers of final goods in nontradable industries remain sheltered from foreign competition, and Equations (7)-(8) apply to all regions and not only to the lowest-cost ones. Thus, Equation (44) no longer applies to the producers of final goods in nontradable industries (Equation (43) still stands as a definition, though).

Market clearing conditions are also affected by the presence of transport costs. While Equations (45)-(46) describing market clearing in regional factor markets still apply, Equation (11) describing market clearing in global markets for final goods applies only to tradable industries. In nontradable industries, Equation (11) must be replaced by analogue conditions imposing market clearing in each regional market:

(85) $P_{c,it} \cdot Q_{c,it} = E_{c,it}$ for all $i \in N_t$ and $c \in C$

This completes the formal description of the model. For any admissible set of capital stocks, i.e. $K_{c,0}$ for all $c \in C$; sequences for the vectors of savings, human capital and industry productivities, i.e. $S_{c,t}$, $H_{c,t}$, and $A_{c,it}$ for all $c \in C$ and for all $i \in I$; and a sequence for the set N_t (or T_t); an equilibrium of the world economy after globalization consists of a sequence of prices and quantities such that the equations listed above hold at all dates and states of nature. Although there might be multiple geographical patterns of production that are consistent with world equilibrium, the assumptions made ensure that prices and world aggregates are uniquely determined.

The best way to start the analysis is by asking again whether the assumed trade restrictions matter at all. That is, to ask whether restricting factor mobility and

now goods trade impede the world to achieve the level of efficiency of the integrated economy. Re-define the set FPE_t now as follows:

$$
\begin{aligned}
FPE_t \equiv \Bigg\{ d_t \in D_t \Bigg| \; & \exists x_{c,it}(m) \geq 0,\ x^F_{c,it} \geq 0 \text{ with } \sum_{c \in C} x_{c,it}(m) = 1,\ \sum_{c \in C} x^F_{c,it} = 1 \text{ and} \\
& x_{c,it} = (1-\beta_i) \cdot x^F_{c,it} + \frac{\beta_i}{M_{it}} \cdot \int_0^{M_{it}} x_{c,it}(m) \cdot dm; \text{ such that :} \\
& \text{(R1) } x_{c,it} = 0 \text{ if } Z_{c,it} < \max_{c \in C} \{Z_{it}\}; \\
& \text{(R2) } H_{c,t} = \sum_{i \in I} x_{c,it} \cdot H_{it} \text{ and } K_{c,t} = \sum_{i \in I} x_{c,it} \cdot K_{it}; \\
& \text{(R3) } x_{c,it}(m) \in \{0,1\} \text{ for all } m \in [0, M_{it}] \text{ and } i \in I; \text{ and} \\
& \text{(R4) } x^F_{c,it} \geq (1-\beta_i) \cdot \frac{Y^{IE}_{c,t}}{Y^{IE}_t} \text{ if } i \in N_t \Bigg\}
\end{aligned}
\tag{86}
$$

Comparing Definitions (36) and (86), we observe that the latter contains an additional requirement: each region should be able to produce all the final goods used for its own consumption and investment in nontradable industries. This additional restriction is a direct consequence of transport costs. The presence of this additional restriction reduces the size of FPE_t. In fact, it is now even possible that $FPE_t = \varnothing$. For instance, assume regional differences in industry productivities are such that there exists no region that has the highest possible productivity in all nontradable industries simultaneously. Then, it is not possible to replicate the integrated economy.[63]

If $d_t \in FPE_t$, factor prices are equalized across regions and the world economy operates with the same efficiency as the integrated economy despite factor immobility and goods market segmentation. In this case, the world economy

[63] That one or more regions with the highest possible productivity in all nontradable industries exist is a necessary but not sufficient condition for $FPE_t \neq \varnothing$. Since factor-price equalization requires that all factors be located in these regions, it is also necessary that at least one of these regions have the highest possible productivity for each tradable industry.

behaves exactly as the world of factor-price equalization of section 2.2.[64] If $d_t \notin FPE_t$, the world economy cannot operate at the same level efficiency as the integrated economy. As a result, either market-based factor prices, or real factor prices, or both differ across regions. But even in this case the behavior of the world economy does not depart much from what we observed in the worlds of section 2. To see this, define $H_{c,t}^T$ and $K_{c,t}^T$ as the factor endowments devoted to the production of tradable goods, i.e. all intermediate inputs and the final goods of tradable industries. Straightforward algebra shows that:[65]

$$(87)\quad H_{c,t}^T = \max\left\{0, H_{c,t}\cdot\left(1-\sum_{i\in N_t}(1-\beta_i)\cdot(1-\alpha_i)\cdot\sigma_i\right) - K_{c,t}\cdot\left(\frac{w_{c,t}}{r_{c,t}}\right)^{-1}\cdot\sum_{i\in N_t}(1-\beta_i)\cdot(1-\alpha_i)\cdot\sigma_i\right\} \text{ for all } c\in C$$

$$(88)\quad K_{c,t}^T = \max\left\{0, K_{c,t}\cdot\left(1-\sum_{i\in N_t}(1-\beta_i)\cdot\alpha_i\cdot\sigma_i\right) - H_{c,t}\cdot\frac{w_{c,t}}{r_{c,t}}\cdot\sum_{i\in N_t}(1-\beta_i)\cdot\alpha_i\cdot\sigma_i\right\} \text{ for all } c\in C$$

Equations (87)-(88) show the factor supplies that are left after subtracting from aggregate factor supplies the factors used in the production of final goods in nontradable industries. In the special case in which $N_t = \varnothing$, these factor supplies equal the aggregate factor supplies and are independent of factor prices. But in the general case, these factor supplies depend on factor prices in the usual way. The higher is the wage-rental, the lower is the human to physical capital ratio used for the production of final goods in nontradable industries and, as a result, the higher is the relative supply of human to physical capital that is left after production of final goods in nontradable industries.

[64] Even the price levels would be equalized across regions, i.e. $P_{c,t}=1$ for all $c\in C$. Note however that there is less indeterminacy regarding the patterns of production and trade, since nontradable final goods must now be produced in the same region where they are used for consumption or investment.

[65] To see this, note that the shares of human and physical capital devoted to producing the final good of the i^{th} nontradable industry are $(1-\beta_i)\cdot(1-\alpha_i)$ and $(1-\beta_i)\cdot\alpha_i$. Add over industries and note that the share of spending in the i^{th} industry is $\sigma_i\cdot Y_{c,t}$.

With Equations (87)-(88) at hand, it is straightforward to see that all the results in sections 2.4 and 2.5 regarding incomes and factor prices still go through in the presence of nontradable final goods. Take, for instance, Example 2.4.1. Equations (47)-(48) must be rewritten as follows:

$$(89) \qquad P_{it} = \sigma_i \cdot \frac{\prod_{i' \in T_t} \left(\sum_{c \in C} Q_{c,i't} \right)^{\frac{\sigma_i}{\sum_{i' \in T_t} \sigma_{i'}}}}{\sum_{c \in C} Q_{c,it}} \qquad \text{for all } i \in T_t$$

$$(90) \qquad Y_t = \prod_{i \in T_t} \left(\sum_{c \in C} Q_{c,it} \right)^{\frac{\sigma_i}{\sum_{i' \in T_t} \sigma_{i'}}}$$

while Equations (49)-(52) still apply provided that we write $H^T_{c,t}$ and $K^T_{c,t}$ instead of $H_{c,t}$ and $K_{c,t}$, in Equations (49)-(50). Factor prices and the pattern of trade are determined by these modified versions of Equations (47)-(52) together with Equations (87)-(88). Since factor supplies are well behaved, a brief analysis of this system reveals that all the discussion of the properties of the world income distribution and its dynamics after Equations (58)-(59) still goes through. In fact, all the results and intuitions developed in the examples of sections 2.4 and 2.5 still apply after we remove the assumption that $N_t = \varnothing$.

The major difference between the world of this sub-section and the one in section 2 is that there is a discrepancy between market-based and real incomes and factor prices. To see this, we need to compute regional price levels. Equations (5)-(7) and (83) imply that:

$$(91) \qquad P_{c,t} = \prod_{i \in N_t} \left\{ \frac{1}{\sigma_i} \cdot \left[\frac{1}{Z_{c,it}} \cdot \left(\frac{w_{c,t}}{1-\alpha_i} \right)^{1-\alpha_i} \cdot \left(\frac{r_{c,t}}{\alpha_i} \right)^{\alpha_i} \right]^{1-\beta_i} \cdot \left[\int_0^{M_{it}} p_{it}(m)^{1-\varepsilon_i} \cdot dm \right]^{\frac{\beta_i}{1-\varepsilon_i}} \right\}^{\sigma_i}$$

Since all regions face the same input prices, Equation (91) shows that, ceteris paribus, the price level is high in regions that have high factor prices and low productivity in nontradable industries. This relationship is the first piece of a theory of the price level. The second piece is a relationship between factor prices, factor endowments and industry productivities. The following examples show how to obtain this additional relationship.

Example 3.1.1: Consider a world economy with H- and K-industries, $I^H \cup I^K = I$ and $I^H \cap I^K = \varnothing$. Assume $\alpha_i=\alpha_H$ and $\max_{c\in C}\{Z_{c,it}\} = Z_{Ht}$ if $i\in I^H$, $\alpha_i=\alpha_K$ and $\max_{c\in C}\{Z_{c,it}\} = Z_{Kt}$ if $i\in I^K$, with $\alpha_H \leq \alpha_K$. For simplicity, assume also that $\varepsilon_i=\varepsilon$ and $\beta_i=\beta$ for all $i\in I$. As in section 2.4, we assume that requirement R1 is not binding.[66] The only difference between this world and the one in Example 2.4.1 is the presence of nontradable industries, i.e. $N_t \neq \varnothing$.

Let P_{Ht} and P_{Kt} be the prices of final goods in tradable H- and K-industries. If a region is internationally competitive in tradable H-industries, then the price of final goods of its nontradable H-industries is also P_{Ht}.[67] If a region is not competitive internationally, then the price of final goods in its nontradable H-industries exceeds P_{Ht}. In fact, it follows from Equations (5) and (51)-(52) that the price of the final goods in nontradable H-industries is $\frac{w_{c,t}^{1-\alpha_H} \cdot r_{c,t}^{\alpha_H}}{f_{Ht}} \cdot P_{Ht} \geq 1$. A parallel argument shows that the price of the final goods in nontradable K-industries is $\frac{w_{c,t}^{1-\alpha_K} \cdot r_{c,t}^{\alpha_K}}{f_{Kt}} \cdot P_{Kt} \geq 1$. It then follows from Equations (83)-(84) that the price level of region c is given by:

[66] Note that this implies that all regions have the same productivity in nontradable industries. That is, $Z_{c,it}=Z_{Ht}$ if $i\in N_t \cap I^H$ and $Z_{c,it}=Z_{Kt}$ if $i\in N_t \cap I^K$ for all $c\in C$.

[67] This follows because the technology to produce final goods is the same for all H-industries, and also because the number of input varieties of H-industries does not depend on whether the industry is tradable or nontradable.

(92) $$P_{c,t} = \left[\frac{w_{c,t}^{1-\alpha_H} \cdot r_{c,t}^{\alpha_H}}{f_{Ht}} \cdot \frac{P_{Ht}}{\sigma_H} \right]^{\sum_{i \in N_t \cap I^H} \sigma_i} \cdot \left[\frac{w_{c,t}^{1-\alpha_K} \cdot r_{c,t}^{\alpha_K}}{f_{Kt}} \cdot \frac{P_{Kt}}{\sigma_K} \right]^{\sum_{i \in N_t \cap I^K} \sigma_i} \quad \text{for all } c \in C$$

As in Example 2.4.1, regions with intermediate factor proportions have diversified production structures while regions with extreme factor proportions have specialized production structures. The sets C_{Mt} , C_{Kt} and C_{Ht} are still defined by Equations (53)-(55) provided we write $H_{c,t}^T$ and $K_{c,t}^T$ instead of $H_{c,t}$ and $K_{c,t}$. It follows from Equations (51), (52) and (92) that $P_{c,t} = \left(\frac{P_{Ht}}{\sigma_H} \right)^{\sum_{i \in N_t \cap I^H} \sigma_i} \cdot \left(\frac{P_{Kt}}{\sigma_K} \right)^{\sum_{i \in N_t \cap I^K} \sigma_i}$ if $c \in C_{Mt}$ and $P_{c,t} \geq \left(\frac{P_{Ht}}{\sigma_H} \right)^{\sum_{i \in N_t \cap I^H} \sigma_i} \cdot \left(\frac{P_{Kt}}{\sigma_K} \right)^{\sum_{i \in N_t \cap I^K} \sigma_i}$ if $c \in C_{Kt} \cup C_{Ht}$. All regions within the cone share the same price level, and this is the lowest in the world. The reason, of course, is that these regions are competitive both in H- and K-industries. Regions outside the cone have different price levels. Moreover, it is possible to show that these price levels increase the farther away the regions are from the cone. The reason is that the farther away from the cone, the less competitive a region is in one of the industry types and the more expensive it is to produce the final goods of the nontradable industries of this type.

Example 3.1.1 provides us with a simple theory of why and how the price level varies across regions. But it is difficult to reconcile this theory with the data. The later show that price levels are positively correlated with income, so that regional differences in real incomes are substantially smaller than regional differences in market-based incomes. To obtain this pattern in the world of Example 3.1.1 would require that poor regions be located inside the cone and rich regions outside of it. Although this is not impossible from a theoretical standpoint, it does not seem a

promising starting point for the construction of an empirically successful theory of the price level.

A positive association between incomes and price levels could arise somewhat more naturally in the world of Example 2.4.2 once we remove the assumption that $N_t = \varnothing$. For instance, if nontradable industries tend to be more human-capital intensive than tradable industries the price level would be high in regions that belong to C_{Kt}; intermediate in regions that belong to C_{Mt}; and low in regions that belong to C_{Ht}. Assume then that most of the variation in income levels is due to differences in savings rates, so that rich regions are those that have low human to physical capital ratios. This does not seem implausible, since most nontradable industries tend to be in the service sector and this sector tends to use a higher human to physical capital ratio than other sectors.

More generally, in the worlds of sub-section 2.4 the correlation between income and price levels is positive or negative depending on how factor proportions vary with income and the factor intensities of nontradable industries relative to tradable ones. The central observation is that price levels should be high in regions that have factor proportions that are inadequate to produce nontradable goods. Building an empirically successful theory of the price level around this notion seems promising, although it remains to be done. Most of the existing research on the price level has focused instead on the role of regional differences in industry productivities. The next example presents a world where these differences generate a positive association between income and the price level.

Example 3.1.2: Consider a world where $Z_{c,it} = \pi_{c,Ht}^{1-\alpha_i} \cdot \pi_{c,Kt}^{\alpha_i}$ for all $i \in T_t$ and $c \in C$, and $Z_{c,it}=1$ for all $i \in N_t$, with $\sum_{c \in C} \pi_{c,Ht} \cdot \frac{H_{c,t}}{H_t} = 1$ and $\sum_{c \in C} \pi_{c,Kt} \cdot \frac{K_{c,t}}{K_t} = 1$. The crucial feature

of this example is that productivity differences exist only in tradable industries.[68] This world economy is akin to that in Example 2.5.1. For instance, assume that there are H- and K-industries as in Examples 2.4.1 and 3.1.1. Then, we have that:

$$(93) \quad P_{c,t} = \left[\frac{(\pi_{c,Ht} \cdot \hat{w}_{c,t})^{1-\alpha_H} \cdot (\pi_{c,Kt} \cdot \hat{r}_{c,t})^{\alpha_H}}{\hat{f}_{Ht}} \cdot \frac{P_{Ht}}{\sigma_H} \right]^{\sum_{i \in N_t \cap I^H} \sigma_i} \cdot \left[\frac{(\pi_{c,Ht} \cdot \hat{w}_{c,t})^{1-\alpha_K} \cdot (\pi_{c,Kt} \cdot \hat{r}_{c,t})^{\alpha_K}}{\hat{f}_{Kt}} \cdot \frac{P_{Ht}}{\sigma_H} \right]^{\sum_{i \in N_t \cap I^K} \sigma_i} \quad \text{for all } c \in C$$

where $\hat{f}_{it} = \min_{c \in C} \left\{ \hat{w}_{c,t}^{1-\alpha_i} \cdot \hat{r}_{c,t}^{\alpha_i} \right\}$ for all $i \in I$. Since productivity differences in tradable industries are factor augmenting, regions with higher productivities have higher factor prices. Since there are no productivity differences in nontradable industries, regions with higher factor prices have a higher price level. Note that now a region inside the cone with high productivity in the tradable industries could have a higher price level than a region outside the cone with low productivity in the tradable industries.

In the world of this example, the price level is determined by a combination of two elements: how adequate are the region's factor proportions to produce in the nontradable industries; and how high is the region's productivity in the tradable industries relative to the nontradable ones. In the world of Example 3.1.1, this second force was not present and Equation (93) was reduced to Equation (92). We could also eliminate the first force by assuming that all regions belong to the cone, i.e. by assuming that there is conditional factor-price equalization. In this case, the price level is given by:

$$(94) \quad P_{c,t} = \left[\pi_{c,Ht}^{1-\alpha_H} \cdot \pi_{c,Kt}^{\alpha_H} \cdot \frac{P_{Ht}}{\sigma_H} \right]^{\sum_{i \in N_t \cap I^H} \sigma_i} \cdot \left[\pi_{c,Ht}^{1-\alpha_K} \cdot \pi_{c,Kt}^{\alpha_K} \cdot \frac{P_{Kt}}{\sigma_K} \right]^{\sum_{i \in N_t \cap I^K} \sigma_i} \quad \text{for all } c \in C$$

[68] This assumption makes sense because nontradable industries consist mostly of services, and in the real world productivity differences in services seem small relative to productivity differences in agriculture or manufacturing.

In Equation (94) the only determinant of the price level is the level of productivity in the tradable industries. This special case is known as the Balassa-Samuelson hypothesis of why the price level is positively correlated with income. Higher productivity in the tradable industries is what makes regions both rich and expensive.

In addition to providing a theory of the price level, the world of this section is also useful because it allows us to study a smoother and more realistic version of the globalization process, i.e. a gradual reduction in the size of N_t. This is not only important for quantitative applications of the theory, but it also leads to new insights regarding the effects of globalization on welfare. The next example shows this.

Example 3.1.3: Consider a world economy with H- and K-industries, such that $I^H \cup I^K = I$ and $I^H \cap I^K = \varnothing$. Assume α_i=0 if $i \in I^H$, and α_i=1 if $i \in I^K$; and β_i=0 for all $i \in I$. Within each type there are "advanced" and "backward" industries. A-regions have the highest possible productivity in all industries, regardless of whether they are "advanced" or "backward". B-regions have the highest possible productivity only in "backward" industries. Up this point all the assumptions are as in Example 2.1.2, except that industry factor intensities are more extreme. Assume next that initially some industries are nontradable, i.e. $N_t \neq \varnothing$; and consider a small step in the globalization process: some "advanced" H-industries become tradable, i.e. some elements of the set $N_t \cap I^H$ move into the set $T_t \cap I^H$. What is the effect of this partial reduction in transport costs on regional incomes?

The reduction of transport costs leads to structural transformation: A-regions reduce their production in "backward" H-industries and increase their production in "advanced" H-industries, while B-regions do the opposite.[69] This increases efficiency

[69] Given the extreme assumptions on industry factor intensities, we know that the distribution of production in K-industries will not be affected.

and raises the combined world production of H-industries, lowering the price of their products and therefore wages all over the world. Therefore, a partial reduction of transport costs has two effects: an increase in efficiency that lowers prices and benefits all regions, and a change in relative prices that benefits some regions but hurts others. A-regions with a large enough ratio of human to physical capital are worse off as a result of this partial reduction in transport costs.[70] If coupled with an appropriate transfer scheme, partial globalization still constitutes a Pareto improvement for the world economy. But now this transfer scheme might require inter-regional transfers towards A-regions with large enough human to physical capital ratios.

The world of this sub-section is a simple and yet very useful generalization of the world of section 2. It allows us to study the sources of regional differences in price levels and also permits us to consider smoother versions of the globalization process. Despite this progress, the world of this sub-section fails to capture a central aspect of transport costs because these only affect final goods. When transport costs affect intermediate inputs, they create incentives to agglomerate production in a single location. We study how this works next.

3.2 Agglomeration effects

Consider a world where transport costs apply only to intermediates, and not to final goods. In particular, assume that the costs of transporting inputs are negligible if $i \in T_t$ but prohibitive if $i \in N_t$; while the costs of trading final goods are negligible for all $i \in I$. An implication of this last assumption is that the price level is the same in all regions and market-based and PPP-adjusted incomes coincide. But this

[70] How is it possible that a region have negative gains from globalization? Since relative prices have changed, the region's trade opportunities have changed also and it might no longer be possible to achieve the level of income and welfare that the region enjoyed before the reduction of transport costs in the H-industries.

does not mean that we are back to the worlds of section 2. The inability of trading intermediate inputs creates an incentive to concentrate all the production of an industry in a single region. Only in this way, production of final goods can fully take advantage from the benefits of specialization. This force towards the agglomeration of economic activity has profound effects on the world income distribution and its dynamics.

The formal description of the model is quite similar to that of section 2.3. Regional analogues to Equations (1)-(3) apply. Since all regions share spending patterns and face the same final goods prices, the price of consumption and investment is the same for all, and we keep Equation (4) as the numeraire rule. Equations (5)-(6) describing technology apply to all regions, with the corresponding factor prices and industry productivities. The only difference with the model of section 2.2 is that, even after globalization, producers of intermediate inputs in nontradable industries remain sheltered from foreign competition. As a result, in these industries Equations (9)-(10) apply to producers of intermediates in all regions and not only to the lowest-cost ones. Also Equation (8) applies to each region separately since only the demand from local producers of final goods matters for the producers of intermediate inputs. Thus, Equation (44) no longer applies to the producers of intermediate inputs in nontradable industries, and Equation (43) must be modified as follows:

$$(95)\quad I_{c,t} \equiv \left\{ i \in I \;\middle|\; c \in \underset{c' \in C}{\operatorname{argmin}} \left\{ \left[\frac{1}{Z_{c,it}} \cdot \left(\frac{w_{c,t}}{1-\alpha_i} \right)^{1-\alpha_i} \cdot \left(\frac{r_{c,t}}{\alpha_i} \right)^{\alpha_i} \right]^{1-\beta_i} \cdot \left[\int_0^{M_{c,it}} p_{c,it}(m)^{1-\varepsilon_i} \cdot dm \right]^{\frac{\beta_i}{1-\varepsilon_i}} \right\} \right\} \quad \text{for all } c \in C$$

Equation (95) simply recognizes that the number of intermediate inputs available and their prices can vary across regions.[71] Finally, the market clearing conditions in Equations (11), (45) and (46) apply.

This completes the formal description of the model. For any admissible set of capital stocks, i.e. $K_{c,0}$ for all $c \in C$, and sequences for the vectors of savings, human capital and industry productivities, i.e. $S_{c,t}$, $H_{c,t}$, and $A_{c,it}$ for all $c \in C$ and for all $i \in I$; and a sequence for the set N_t (or T_t); an equilibrium of the world economy after globalization consists of a sequence of prices and quantities such that the equations listed above hold at all dates and states of nature. Like the other worlds we have studied up to this point, there might be multiple geographical patterns of production that are consistent with world equilibrium. Unlike the worlds we have studied up to this point however, there might also be multiple prices and world aggregates that are consistent with world equilibrium. This is, in fact, the most prominent feature of this world.

As usual, we start the analysis by defining the set of factor distributions that allow the world economy to replicate the integrated economy. This set is now as follows:

$$
\begin{aligned}
FPE_t \equiv \Bigg\{ d_t \in D_t \;\Big|\; & \exists x_{c,it}(m) \geq 0,\ x^F_{c,it} \geq 0 \text{ with } \sum_{c \in C} x_{c,it}(m) = 1,\ \sum_{c \in C} x^F_{c,it} = 1 \text{ and} \\
& x_{c,it} = (1-\beta_i) \cdot x^F_{c,it} + \frac{\beta_i}{M_{it}} \cdot \int_0^{M_{it}} x_{c,it}(m) \cdot dm; \text{ such that:} \\
& \text{(R1) } x_{c,it} = 0 \text{ if } Z_{c,it} < \max_{c \in C}\{Z_{it}\}; \\
& \text{(R2) } H_{c,t} = \sum_{i \in I} x_{c,it} \cdot H_{it} \text{ and } K_{c,t} = \sum_{i \in I} x_{c,it} \cdot K_{it}; \text{ and} \\
& \text{(R3) } x_{c,it} \in \{0,1\}\ i \in I \Bigg\}
\end{aligned}
\tag{96}
$$

[71] Equation (95) assumes that regions always produce intermediates with the lowest indices. This simplifies notation a bit and carries no loss of generality.

When comparing this set to those in Definitions (36) and (86), we observe that requirement (R3) is much stronger now. While Definitions (36) and (86) only required that the entire production of each intermediate were located in a single region, Definition (96) requires that the entire production of each industry (i.e. all intermediates plus final goods) be located in a single region. This is a direct implication of the assumption that intermediate inputs are nontradable. Naturally, this strengthening of requirement R3 reduces the size of FPE_t.[72] Therefore, this set is always smaller than the set in Definition (36). But it need not be smaller than the set in Definition (86), since requirement R4 no longer applies when final goods are tradable.

Assume that industries are "small" and regions are "large" so that requirement R3 is not binding. Then, it is straightforward to see that the equilibria studied in section 2 still apply. If $d_t \in FPE_t$, there exists an equilibrium in which factor prices are equalized across regions and the world economy operates at the same level of efficiency as the integrated economy despite factor immobility and goods market segmentation. If $d_t \notin FPE_t$, the world economy cannot operate at the same level efficiency as the integrated economy and factor prices differ across regions. All the equilibria analyzed in sub-sections 2.4 and 2.5 are also equilibria for the world of this section, and all the results and intuitions we learned in these sub-sections remain valid without qualification.

There is however a major difference between this world and the ones we studied in section 2. While the equilibria described in section 2 were unique in the worlds analyzed there, they are only one among many in the world of this section. The next example makes this point very clear:

[72] The set FPE_t is never empty, but it is smaller than the set of all the factor distributions that are equilibria of the integrated economy. The reason is that some of these equilibria split industries across regions.

Example 3.2.1: Consider a world where all industries are nontradable, i.e. $N_t = I$. Then, any collection of sets $I_{c,t}$ (with $I_{c,t} \neq \varnothing$ for all $c \in C$) that constitutes a partition of I is part of an equilibrium of the world economy.[73] This follows immediately from Equations (5) and (8), which now apply to each region, and Equation (95). Equation (5) shows that the cost of production of final good producers in a given region depends on the number of available inputs. But Equation (8) shows the number of inputs produced in a given region depends on the demand by local producers of final goods.

This world economy exhibits a very strong form of agglomeration effects, as a result of backward linkages in production.[74] If there are no input producers in a region, the cost of producing final goods is infinity and no final goods producer will choose to locate in the region. But if there are no final goods producers in a region, there is no demand for inputs and no input producer will choose to locate in the region. In this world economy, these forces for agglomeration are so strong that they dwarf comparative advantage. It is possible that a given industry locates in a region offering cheap factors and high productivity, but it is also possible that it ends up locating in region offering expensive factors and low productivity.

The world income distribution can be written as follows:

(97) $$\frac{Y_{c,t}}{Y_t} = \sum_{i \in I_{c,t}} \sigma_i \qquad \text{for all } c \in C$$

[73] This world economy also has equilibria in which industries are split across regions. In these equilibria, all the regions that host a given industry have the same costs of producing the final goods but possibly different numbers and prices of inputs.

[74] Helpman and Krugman [1985] define a backward linkage as a situation in which a final good producer demands many inputs; and a forward linkage as a situation in which many final good producers demand the same input.

Equation (97) is formally very similar to Equation (76). Remember that the latter described the world income distribution in Example 2.5.2 where differences in industry productivities were so strong so as to single-handedly determine comparative advantage. The formal similarity between these two worlds follows because both exhibit an extreme form of specialization. The difference, of course, is the underlying force that determines this specialization. While in Example 2.5.2 regions specialize in a given industry because of their high productivity, in Example 3.2.1 regions specialize in a given industry only because of luck. While in Example 2.5.2 the shape and evolution of the world income distribution reflects only the distribution of industry productivities, in Example 3.2.1 it reflects only randomness.[75]

Example 3.2.1 is extreme because it assumes all industries are nontradable. Assume instead that $T_t \neq \varnothing$, and let $I_{c,t}^{N_t} = N_t \cap I_{c,t}$. As a result of agglomeration effects, any collection of sets $I_{c,t}^{N_t}$ (with $I_{c,t}^{N_t} \neq \varnothing$ for all $c \in C$) that constitutes a partition of N_t is an equilibrium of the world economy. Let again $H_{c,t}^T$ and $K_{c,t}^T$ be the factor endowments used in the production of tradable goods, i.e. all final goods and the intermediate inputs of tradable industries. It follows that:[76]

$$(98) \qquad H_{c,t}^T = \max\left\{0, H_{c,t} - \sum_{i \in I_{c,t}^{N_t}} (1-\alpha_i) \cdot \sigma_i \cdot \frac{Y_t}{w_{c,t}}\right\} \quad \text{for all } c \in C$$

$$(99) \qquad K_{c,t}^T = \max\left\{0, K_{c,t} - \sum_{i \in I_{c,t}^{N_t}} \alpha_i \cdot \sigma_i \cdot \frac{Y_t}{r_{c,t}}\right\} \quad \text{for all } c \in C$$

[75] Given our assumption of full depreciation of inputs, nothing prevents the pattern of production to shift randomly from generation to generation. This model therefore is consistent with any dynamics for the world income distribution. If inputs depreciated slowly, initial randomness would persist for some time.
[76] Here, I am assuming that industries do not split across regions. As mentioned in an earlier footnote, this is possible too.

Equations (98)-(99) show the factor supplies that are left after subtracting from aggregate factor supplies the factors used in nontradable industries. These Equations are analogous to Equations (87)-(88) of sub-section 3.1. One can use Equations (98)-(99) and a given collection of sets $I_{c,t}^{N}$ to generalize the theory of sections 2.4 and 2.5. For instance, in Example 2.4.1 Equations (47)-(52) still apply provided that we write $H_{c,t}^{T}$ and $K_{c,t}^{T}$ instead of $H_{c,t}$ and $K_{c,t}$.

The effects of this generalization of the theory are hard to assess given the multiplicity of equilibria and the inherent difficulty of finding a "respectable" selection criteria.[77] It is always possible to find perverse equilibria in which regions specialize in the "wrong" industries, i.e. industries in which they do not have comparative advantage. Naturally, all the equilibria of section 2 in which regions specialize in the industries in which they have comparative advantage still apply if requirement R3 is not binding (as we have assumed so far). But there is no compelling reason to choose them over some of the alternatives. Moreover, if requirement R3 is violated or is binding, the equilibria studied in section 2 no longer apply to this world economy. The following example, inspired by Krugman and Venables [1995], relaxes the assumption that industries are "small" and clearly illustrates this point:

Example 3.2.2: Consider a world with two industries I={A,M} and two regions C={N,S}. Assume that both industries have the same factor intensities, i.e. $\alpha_i=\alpha$ for all $i \in I$; but different sizes $\sigma_A<0.5<\sigma_M$ (remember that $\sigma_A+\sigma_M=1$). Also assume that both regions are identical, i.e. they have the same savings, human capital, industry productivities and initial condition. Assume next that the world starts in autarky and globalization proceeds in two stages: in the first one industry A becomes tradable, i.e. N_t={M} for $0 \leq t<T$; and in the second stage also industry M becomes tradable, i.e.

[77] Matsuyama [1991], Krugman [1991] and Fukao and Benabou [1993] study some interesting ways of resolving this indeterminacy.

$N_t = \varnothing$ for $t \geq T$. In the world of autarky, both regions have the same income and the question that I shall address here is: How does globalization affect the world income distribution?

At date t=0, all transport costs disappear except for those that affect the intermediate inputs of industry M. There are two possible patterns of production and trade that can emerge as a result of this. The first one consists of both regions producing the same they did in autarky and not trading between them. Since both regions would have the same goods and factor prices, there would be no incentive for any producer to deviate from this equilibrium. The second possible pattern of production and trade that can emerge consists of each region specializing in a different industry. For instance, assume N specializes in industry M. The absence of other local producers in industry M means that producers in S have no incentive to produce in industry M. Since spending on industry M is more than half of world spending, factor prices are higher in N and therefore producers in N cannot compete in industry A.[78]

It follows from this discussion that the first stage of globalization generates world inequality and world instability. In the world of autarky, both regions had the same income level and income volatility was driven by volatility in fundamentals, i.e. savings, human capital and industry productivities. Globalization generates divergence in incomes because in the equilibrium with specialization the region that "captures" industry M has higher income than the region that is "stuck" producing in industry A. The world income distribution is determined by Equation (97). One effect of this inequality is faster physical capital accumulation in N than in S. Globalization also generates instability, since the pattern of specialization can now change capriciously just as a result of a change in expectations. At any time the specialization pattern can change to the detriment of N and to the advantage of S.

[78] The assumption that industry M is large is crucial in reducing the number of equilibria to three. If there were many "small" M-industries there would also be additional equilibria that split these industries between regions in many different ways.

This constitutes an additional source of income volatility that goes beyond fundamentals.

At date t=T, transport costs for the intermediate inputs of industry M vanish. Although the pattern of production and trade is not uniquely determined, we know that factor prices and incomes are uniquely determined.[79] Moreover, since we have assumed that both industries have the same factor intensities, the world income distribution is now given by Equation (70). It follows that the second stage of globalization starts a slow process of convergence in incomes that eventually restores equality across regions. Throughout this process, expectations no longer play any role and the only sources of income volatility are fluctuations in fundamentals.

This example features a combination of agglomeration effects and "large" industries that underlies most of the work known as economic geography.[80] This research has focused on explaining how income differences can arise among regions that initially have the same fundamentals. The view of globalization and development that arises from this literature is colorful and suggestive, although it has not been subjected yet to serious empirical analysis.

Not surprisingly, globalization might lead to a Pareto-inferior outcome in the world of this section. The following example, which is related to Examples 2.1.2 and 3.1.3, shows this:

[79] When $N_t=\varnothing$, we are back to the world of section 2. The reason why the pattern of production is indeterminate is because I have assumed that industry A and M have the same factor intensities. Otherwise we would be in the case of Example 2.1.1.
[80] See Fujita, Krugman and Venables [1999] and Baldwin, Forslid, Martin, Ottaviano and Robert-Nicaud [2003].

Example 3.2.3: Consider a world economy with H- and K-industries, such that $I^H \cup I^K = I$ and $I^H \cap I^K = \varnothing$. Assume α_i=0 if $i \in I^H$, and α_i=1 if $i \in I^K$; and β_i is small (but not zero) for all $i \in I$. Within each type there are "advanced" and "backward" industries. A-regions have the highest possible productivity in all industries, regardless of whether they are "advanced" or "backward". B-regions have the highest possible productivity only in "backward" industries. Assume next that after globalization all industries are non-tradable. This world is just a special case of Example 3.2.1. We know therefore that there is an equilibrium in which A-regions specialize in "backwards" industries while B-regions specialize in "advanced" industries. This equilibrium can be easily shown to deliver equal or less income and welfare than autarky. Since β_i is small for all $i \in I$, the benefits from an increase in market size are negligible. Since the allocation of production worsens relative to autarky, production and income go down as a result of globalization. Therefore, it is not possible to find a transfer scheme that ensures that globalization benefits all.

Although this is real a theoretical possibility, it is not clear yet how seriously should we take the possibility that globalization worsens the world allocation of production and reduces welfare. How important empirically are these agglomeration effects? What is the relative importance of randomness and comparative advantage in determining the pattern of production and trade? The answers to these questions are critical in determining whether the basic policy prescription that simply opening up to trade leads to development really applies or not. In the worlds of this section, opening up to trade can lead to miracles and disasters alike. A miracle is nothing but a lucky region that attracts a large number of industries exhibiting agglomeration effects. A disaster is an unlucky region that cannot do so. Opening up to trade is therefore a gamble. It opens the door for industries to come into the region and enrich it, but it also opens the door for industries to leave the region and impoverish it. Naturally, the temptation to change the odds of this gamble using industrial policies and protectionism might be overwhelming. The prescriptions for

development are therefore easy to spot but not pleasant. This is a world characterized by negative international spillovers and strong temptations to use "beggar-thy-neighbor" policies.

Despite the presence of transport costs, differences in regional market size still play no role in determining the world income distribution in the worlds of this sub-section and the previous one. If intermediate inputs are tradable, all regions use the same set of specialized inputs and enjoy the same level of industry specialization or technology to produce final goods. If final goods are tradable, industries concentrate their production in one or few regions and all regions buy their final goods at the same prices. The ability to trade intermediates and/or final goods therefore implies that regional differences in market size cannot be a source of regional differences in incomes. We next turn to a world that features some industries in which neither intermediates nor final goods can be traded. This brings back market size effects as a determinant of the world income distribution.

3.3 The role of local markets

We turn next to a world in which the costs of trading intermediate inputs and final goods are prohibitive if $i \in N_t$, but negligible if $i \in T_t$. As in all the worlds considered in this chapter, the benefits of developing specialized inputs depend on the size of the industry's market. For tradable industries, this market is the world economy. For nontradable industries, this market is the region. As a result, regional differences in market size will be translated into regional differences in the degree of specialization or technology of nontradable industries. [81]

[81] There is little empirical evidence that regional differences in market size are an important determinant of income differences. When one interprets the data from the vantage point of the world of autarky, this observation implies that market size effects are weak and sustained growth is not possible. This has led many researchers to spend a substantial effort in developing autarky models where sustained growth is possible without market size effects. Somewhat ironically, once one takes a world equilibrium view of the growth process what requires a substantial effort is to develop models where regional differences in market size do affect the world income distribution.

Formally, this model is very similar to the one in sub-section 3.1. Equations (81)-(82) describe investment and consumption, while Equations (83)-(84) still provide the numeraire rule and the price level. Naturally, Equation (3) describing spending patterns still applies to all regions, and Equations (5)-(6) describing technology apply to all regions, with the corresponding factor prices and industry productivities. The only difference with the model of section 3.1 is when Equations Equations (7)-(10) describing pricing policies, input demands and the free-entry condition apply. For tradable industries, these Equations apply only to those regions where the lowest-cost producers are located. For nontradable industries, these Equations apply to all regions and not only to the lowest-cost ones. Thus, Equation (44) no longer applies to producers in nontradable industries, and Equation (93) must be replaced by Equation (95). Market clearing conditions are also the same as in the model of section 3.1, and consist of Equations (45)-(46) describing market clearing in regional factor markets, Equation (11) describing market clearing in global markets for tradable industries, and Equation (85) describing market clearing in regional markets for nontradable industries.

This completes the formal description of the model. For any admissible set of capital stocks, i.e. $K_{c,0}$ for all $c \in C$; sequences for the vectors of savings, human capital and industry productivities, i.e. $S_{c,t}$, $H_{c,t}$, and $A_{c,it}$ for all $c \in C$ and for all $i \in I$; and a sequence for the set N_t (or T_t); an equilibrium of the world economy after globalization consists of a sequence of prices and quantities such that the equations listed above hold in all dates and states of nature. Like other worlds we have studied up to now, there might be multiple geographical patterns of production that are consistent with world equilibrium. But unlike the world of the previous sub-section (and like the worlds of section 2 and sub-section 3.1), prices and world aggregates are uniquely determined.

In this world economy, the set FPE_t is empty. Since intermediate inputs that are produced in a region cannot be used in another region, the world economy cannot reach the level of efficiency of the integrated economy.[82] Despite this, it is relatively straightforward to analyze this world. Define again $H_{c,t}^T$ and $K_{c,t}^T$ as the factor endowments devoted to the production of tradable goods, i.e. all intermediate inputs and final goods of tradable industries. Straightforward algebra shows that: [83]

$$(100)\quad H_{c,t}^T = \max\left\{0, H_{c,t}\cdot\left(1-\sum_{i\in N_t}(1-\alpha_i)\cdot\sigma_i\right) - K_{c,t}\cdot\left(\frac{w_{c,t}}{r_{c,t}}\right)^{-1}\cdot\sum_{i\in N_t}(1-\alpha_i)\cdot\sigma_i\right\} \quad \text{for all } c\in C$$

$$(101)\quad K_{c,t}^T = \max\left\{0, K_{c,t}\cdot\left(1-\sum_{i\in N_t}\alpha_i\cdot\sigma_i\right) - H_{c,t}\cdot\frac{w_{c,t}}{r_{c,t}}\cdot\sum_{i\in N_t}\cdot\alpha_i\cdot\sigma_i\right\} \quad \text{for all } c\in C$$

Since factor supplies are well behaved, all the results in sections 2.4 and 2.5 regarding market-based incomes and factor prices still go through in the presence of nontradable industries. As in sub-section 3.1, the only important difference between the world of this sub-section and the one in section 2 is that there is a discrepancy between market-based and real incomes and factor prices. In particular, we can write the price level of region c as follows:

$$(102)\quad P_{c,t} = \prod_{i\in N_t}\left\{\frac{1}{\sigma_i}\cdot\left[\frac{1}{Z_{c,it}}\cdot\left(\frac{w_{c,t}}{1-\alpha_i}\right)^{1-\alpha_i}\cdot\left(\frac{r_{c,t}}{\alpha_i}\right)^{\alpha_i}\right]^{1-\beta_i}\cdot\left[\int_0^{M_{c,it}} p_{c,it}(m)^{1-\varepsilon_i}\cdot dm\right]^{\frac{\beta_i}{1-\varepsilon_i}}\right\}^{\sigma_i} \quad \text{for all } c\in C$$

[82] The set FPE_t might be non-empty in the limiting case where $\beta_i\to 0$ (or $\varepsilon_i\to\infty$) for all $i\in I$. But that limiting case brings us to the world of sub-section 3.1.

[83] To see this, note that the shares of human and physical capital devoted to producing the final good of the i^{th} nontradable industry are now $(1-\alpha_i)$ and α_i. Add over industries and note that the share of spending in the i^{th} industry is $\sigma_i\cdot Y_{c,t}$.

The only difference between this Equation and Equation (91) is that the number and price of intermediate inputs varies across regions. Using Equations (6) and (10), we can transform Equation (103) into the following:

$$(103)\quad P_{c,t} = \prod_{i \in N_t} \left\{ \frac{1}{\sigma_i^{\mu_i} \cdot A_{c,it}} \cdot \left(\frac{w_{c,t}}{1-\alpha_i} \right)^{(1-\alpha_i)\cdot\mu_i} \cdot \left(\frac{r_{c,t}}{\alpha_i} \right)^{\alpha_i \cdot \mu_i} \cdot Y_{c,t}^{1-\mu_i} \right\}^{\sigma_i}$$

Basically, this model brings another element to the theory of the price level. To the extent that nontradable industries exhibit increasing returns, regions with larger markets have lower price levels and higher real incomes.

It is straightforward to re-do some of the previous examples in the context of this world. But I shall not do this. The picture that this world generates is clear and unappealing form an empirical standpoint: regional differences in market size are reflected in regional differences in price levels. Ceteris paribus, larger local markets do not lead to higher market-based incomes and factor prices. But they do lead to lower price levels and, as a result, to higher real incomes and factor prices. This is clearly counterfactual.

4. Final remarks

This chapter has developed a unified and yet tractable framework that integrates many key insights of the fields of international trade and economic growth. Its distinguishing feature is that it provides a global view of the growth process, that is, a view that treats different regions of the world as parts of a single whole. This framework incorporates the standard idea that economic growth in the world economy is determined by a tension between diminishing returns and market size effects to capital accumulation. A substantial effort has been made to show how

trade frictions of various sorts determine the shape of the world income distribution and its dynamics.

Despite the length of this chapter, some important topics have been left out. The first and most glaring omission is asset trade. This type of trade allows the world economy to redirect its investment towards regions that offer the highest risk-adjusted return.[84] To the extent that patterns of trade are determined by comparative advantage, these are the regions where capital is scarce and productive and this raises efficiency in the world economy. To the extent that patterns of trade are determined by luck, asset trade magnifies the effect of this randomness and this could either raise or lower the efficiency of the world economy. If this were all there is to asset trade, it would not be too difficult to add to this chapter a section on asset trade in which we endow the world economy with a complete set of asset markets. But asset trade does not seem to work as the standard theory of complete markets would suggest. Empirically asset trade seems both much smaller and much more volatile than it would be warranted by its fundamentals, i.e. savings, human capital and industry productivities. To understand these aspects of asset trade it seems necessary to incorporate to the theory features such as sovereign risk, asymmetric information and asset bubbles. Although this is a very important task, it would require another chapter of this magnitude and must therefore be left for future work.[85]

A second important omission of this chapter is government policy. A central aspect of globalization so far has been its imbalanced nature. While economic integration has proceeded at a relatively fast pace, political integration is advancing at a slower pace or not advancing at all. The world economy today features global (or semi-global) markets but local governments. In this context, globalization can lead to a decline in growth and income through a reduction in the quality of policies.

[84] Naturally, asset trade also allows for a better risk sharing and this raises welfare. Better risk sharing might also increase investment and growth. See Obstfeld [1994].
[85] Among the many papers that study the behavior of financial markets in world equilibrium models, see Gertler and Rogoff [1990], Acemoglu and Zilibotti [1997], Ventura [2002], Matsuyama [2004], Martin and Rey [2002, 2004], Kraay, Loayza, Servén and Ventura [2005] and Broner and Ventura [2005].

International spillovers eliminate the incentives to adopt good but costly policies. Trade also "bails out" regions with bad policies since they can spare some of their costs by specializing in industries where bad policies have little effects. As a result of these forces, globalization could create a "race to the bottom" in policies that lowers savings, human capital, and industry productivities. And this could potentially mitigate or even reverse the benefits from economic integration.[86] Understanding the circumstances under which this "race to the bottom" can happen and the appropriate policy corrections that are required to allow the world economy to take full advantage of globalization is another important task. But this task would also require another chapter of this magnitude and cannot be undertaken here.

At first sight, factor movements might seem a third important omission. But I think it is less so. As mentioned in section 2, the notion that physical and human capital is geographically immobile seems a fair description of reality. Moreover, the benefits of factor mobility might be reaped without factors having to move at all. What is really important about factor movements is that they permit factors located in different regions to work together and produce. Advances in telecommunications technology and the standardization of software allow producers around the world to combine physical and human capital located in different regions in a single production process. We can always think of this situation as one in which the production process has been broken down into intermediate inputs. An increased ability to combine factors located in different regions could therefore be modeled as an increase in the tradability of intermediate inputs, or as an increase in the share of intermediate inputs, or as the development of additional inputs with more extreme factor intensities. All of these possibilities could be (and some have already been) analyzed within the framework developed in this chapter.[87]

---

[86] See Levchenko [2004] for a situation in which globalization leads to a "race to the top" in government policies, though.

[87] An increase in the tradability of inputs corresponds to a gradual increase in T_t in the models of section 3.2 and 3.3. An increase in the share of intermediate inputs corresponds to a gradual increase in β_i, while the development of inputs with more extreme factor intensities corresponds to a gradual change in α_i. I have assumed throughout that industry characteristics are time-invariant only for simplicity. All the formulas in this chapter remain valid if we instead assume that industry characteristics vary, perhaps stochastically, over time.

The goal of this chapter has been to convey a global way of thinking about the growth process. To claim success, you should be persuaded by now that developing and systematically studying world equilibrium models is a necessary condition to gain a true understanding of the growth process. By “true”, I mean the sort of understanding that allows us to frame clear and unambiguous hypotheses about why some countries are richer than others or what are the main forces that drive economic growth in the world economy. To claim success, you should also be convinced by now that much is already known about the structure of world equilibrium models. But you should also be aware that the global view of economic growth that these models reveal is still somewhat fuzzy and blurred. Sharpening this view is a major challenge for growth and trade theorists alike.

5. References

Acemoglu, D. [2003]: “Patterns of Skill Premia”, *Review of Economic Studies*, Vol. 70, pp. 199-230

Acemoglu, D. and Ventura, J. [2002]; “The world Income Distribution,” *Quarterly Journal of Economics* CXVII, pp. 659-694

Acemoglu, D. and Zilibotti, F. [2001]: “Productivity Differences,” *Quarterly Journal of Economics* CXVI, pp. 563-605.

Acemoglu, D. and Zilibotti, F. [1997]: “Was Prometheus Unbound by Chance? Risk, Diversification and Growth”, *Journal of Political Economy* CV, pp.709-751

Ades, A. and Glaeser, E. L. [1999]: “Evidence on Growth, Increasing Returns, and the Extent of the Market,” *Quarterly Journal of Economics*, 114 (3), pp.1025-45.

Alcalá, F. and Ciccone, A. [2003]: “Trade, Extent of the Market, and Economic Growth 1960-96,” mimeo, Universitat Pompeu Fabra

Alcalá, F. and Ciccone, A. [2004]: “Trade and Productivity,” *Quarterly Journal of Economics*, *119* (2), pp. 613-46.

Alesina, A., Spolaore, E., and Wacziarg, R. [2000]: “Economic Integration and Political Disintegration,” *American Economic Review*, *90* (5), pp. 1276-96

Alvarez, F. and Lucas, R. [2004]: “General Equilibirum Analysis of the Eaton-Kourtum model of International Trade,” manuscript, University of Chicago.

Aoki, M. [1986]: “Dynamic Adjustment Behaviour to Anticipated Supply Shocks in a Two-Country Model”, *Economic Journal*, XCVI, pp. 80-100

Arnold, L. G. [2002]: “On the Growth Effect of North-South Trade: the Role of Labor Market Flexibility,” *Journal of International Economics* LVIII, pp. 451-466

Atkeson, A. and Kehoe, P. [1998]: "Paths of Development for Early and Late Bloomers in a Dynamic Hecksher-Ohlin Model", *Federal Reserve Bank of Minneapolis Staff Report* 256

Backus, D., Kehoe, P., and Kydland, F. [1992]:, “International Real Business Cycle”, Journal of Political Economy, C, pp.745-775

Baldwin, R. [1992]: “Measurable Dynamic Gains from Trade” *Journal of Political Economy* C, pp. 162-174

Baldwin, R. and Forslid R. [2000]: “Trade Liberalization and Endogenous Growth. A q-theory Approach,” *Journal of International Economics* L, pp. 497-517

Baldwin, R., Martin, P., and Ottaviano G. I. P. [2001]: “Global Income Divergence, Trade, and Industrialization: The Geography of Growth Take-Offs”, *Journal of Economic Growth*, V pp. 5–37

Baldwin, R., R. Forslid, P. Martin, G. Ottaviano, and F. Robert-Nicaud: "Economic Geography and Public Policy ", Princeton University Press 2003 ISBN 0-691-10275-9

Bardham, P. [1965]: “Optimal Accumulation and International Trade,” *Review of Economics Studies* XXXII, pp. 241-244

Bardham, P. [1965a]: “Equilibrium Growth in the International Economy,” *Quarterly Journal of Economics* LXXIX, pp. 455-464.

Bardham, P. [1966]: “On Factor Accumulation and the Pattern of International Specialization,” *Review of Economics Studies* XXXIII, pp. 39-44.

Barro, R. J. [1991]: “Economic Growth in a Cross Section of Countries”, *Quarterly Journal of Economics*, Vol. 106, No. 2, pp. 407-443

Basu, S. and Weil P. [1998]: “Appropriate Technology and Growth,” *Quarterly Journal of Economics* CXIII, pp. 1025-1053.

Baxter, M. [1992]: “Fiscal Policy, Specialization in the Two-Sector Model: the Return of Ricardo?” *Journal of Political Economy* C, pp. 713-744

Bourguignon, F and Morrison, C. [2002]: “Inequality Among World Citizens: 1820-1992” *American Economic Review*, Vol. 92, No. 4, pp. 727-744

Brems, H. [1956]: “The Foreign Trade Accelerator and the International Transmission of Growth,” *Econometrica* XXIV, pp. 223-238.

Brems, H. [1970]: “ A Growth Model of International Direct Investment,” *American Economic Review* LX, pp. 320-331.

Brezis, S.L., Krugman, P., and Tsiddon D. [1993]: “Leapfrogging in International Competition: A theory of Cycles in National Technological Leadership,” *American Economic Review* LXXXIII, pp. 1211-1219

Broner, F, and Ventura J. [2005]: “Managing Financial Integration,” manuscript, CREI.

Buiter, W.H. [1981]: “ Time Preference and International Lending and Borrowing in an Overlapping-Generations Model” *Journal of Political Economy* LXXIX, pp. 769-797

Caselli, F., Esquivel, G. and Lefort, F. [1996]: “Reopening the Convergence Debate: A New Look at Cross-Country Growth Empirics” *Journal of Economic Growth*

Chang, R. [1990]: “International Coordination of Fiscal Deficits”, *Journal of Monetary Economics*, XXV, pp.347-366

Chui, M., Levine, P., and Pearlman, J. [2001]: “Winners and Losers in a North-South Model of Growth, Innovation and Product Cycles,” *Journal of Development Economics* LXV, pp. 333-365.

Cuñat, A. and Mafezzoli, M. [2004]: “Heckscher-Ohlin Business Cycles”, *Review of Economic Dynamics* 7(3), pp. 555-85.

Cuñat, A. and Mafezzoli, M. [2004a]: “Neoclassical Growth and Commodity Trade”, *Review of Economic Dynamics* 7(3), pp. 707-36.

Davis, D. [1995]: “Intra-industry Trade: A Heckscher-Ohlin-Ricardo Approach,” *Journal of International Economics*, XXXIX, pp. 201-226.

Deardorff, A.V. [2001]: “Rich and poor countries in neoclassical trade and growth”, *Economic Journal*, CXI, pp.277-294.

Devereux, M. and Lapham, B. J. [1994]: “ The Stability of Economic Integration and Endogenous Growth,” *Quarterly Journal of Economics* CIX, pp. 299-305.

Devereux, M. B. and Saito, M. [1997]: “Growth and Risk-Sharing with Incomplete International Assets Markets,” *Journal of International Economics* XLII, pp. 453-481.

Devereux, M. B. and Shi, S. [1991]: “Capital Accumulation and the Current Account in a Two-Country Model,” *Journal of International Economics* L, pp. 1-25.

Dinopoulos, E. and Segerstrom, P. [1999]: “A Schumpeterian Model of Protection and Relative Wages” *American Economic Review*, Vol. 89, No. 3, pp. 450-472
.
Dinopoulos, E. and Syropoulos, C. [1997]: “Tariffs and Schumpeterian Growth,” *Journal of International Economics* XLII, pp. 425-452.

Dixit, A.K., and V. Norman: *Theory of International Trade*, Cambridge University Press, Cambridge, New York, Port Chester, Melbourne, Sydney 1989, ISBN 0521234816

Dollar, D. [1986]: “Technological Innovation, Capital Mobility, and the Product Cycle in North-South Trade,” *American Economic Review* LXXVI, pp. 177-190

Dollar, D. and Kraay, A. [2003]: “Institutions, Trade, and Growth,” *Journal of Monetary Economics*, *50* (1), pp. 133-162

Dornbusch, R., Fischer, S., and Samuelson, P. A. [1977]: “Comparative Advantage, Trade, and Payments in a Ricardian Model with a Continuum of Goods”, *American Economic Review*, Vol. 67 No.5, pp. 823-839

Dornbusch, R., Fischer, S., and Samuelson, P. A. [1980]: “Heckscher-Ohlin Trade Theory with a Continuum of Goods”, *Quarterly Journal of Economics*, Vol. 95, No. 2, pp. 203-224

Eaton, J. and Kortum, S. [2002]: "Technology, Geography, and Trade”, *Econometrica* 70: pp. 1741-1779

Elkan, R. van [1996]: “Catching Up and Slowing Down: Learning and Growth Patterns in an Open Economy,” *Journal of International Economics* XLI, pp. 95-111.

Feenstra, R. C. [1996]: “Trade and Uneven Growth,” *Journal of Development Economics* IL, pp. 229-256

Findlay, R. [1978]: “An ‘Austrian’ Model of International Trade and Interest Rate Equalization”, *Journal of Political Economy*, LXXXVI, pp. 989-1007

Findlay, R. [1980]: “The Terms of Trade and Equilibrium Growth in the World Economy,” *American Economic Review* LXX, pp. 291-299

Findlay, R and Kierzkowski, H. [1983]: “International Trade and Capital: A Simple General Equilibrium Model” *Journal of Political Economy* XCI, pp. 957-978

Fisher, O'N. [1995]: "Growth, Trade, and International Transfers," *Journal of International Economics* IXL, pp. 143-158

Flam, H. and Helpman, E. [1987]: "Vertical Product Differentiation and North-South Trade," *American Economic Review* LXXVII, pp. 810-822

Frankel, J. A. and Romer, D. [1999]: "Does Trade Cause Growth?" *American Economic Review*, Vol. 89 No.3, pp. 379-399

Francois, J. F. [1996]: "Trade, Labour Force Growth and Wages", *Economic Journal*, 439, pp. 1586-1609

Frenkel, J. and Razin, A. [1985]: "Government Spending, Debt, and International Economic Interdependence", *Economics Journal*, XCV, pp.619-636

Frenkel, J. and Razin, A. [1986]: "Fiscal Policies in the World Economy", *Journal of Political Economy*, XCIV, pp. 564-594

Fujita, M. P., Krugman, P. R., and Venables, A. J.: "The Spatial Economy", MIT Press 1999 ISBN 0-262-06204-6

Fukao, K., and R. Benabou [1993]: "History vs. Expectations: A Comment," *Quarterly Journal of Economics*, CVIII, pp. 535-542.

Gale, D. [1971]: "General Equilibrium with Imbalance of Trade," *Journal of International Economics* I, pp. 141-158.

Galor, O. and Polemarchakis, H.M. [1987]: "Intertemporal Equilibrium and the Transfer Paradox", *Review of Economic Studies*, LIV, pp.147-156

Gancia, G. [2003]: "Globalization, Divergence and Stagnation" *IIES Working Paper* #720

Gertler, M. and Rogoff, K. [1990]: "North-South Lending and Endogenous Domestic Capital Market Inefficiencies", *Journal of Monetary Economics*, XXVI, pp. 245-266

Glass, A.J. and Saggi, K. [1998]: "International Technology Transfer and the Technology Gap," *Journal of Development Economics* LV, pp. 369-398

Glass, A.J. and Saggi, K. [2002]: "Intellectual Property Rights and Foreign Direct Investment," *Journal of International Economics* LVI, pp. 387-410

Greenwood, J., Williamson, S.D. [1989]: "International Financial Intermediation and Aggregate Fluctuations under Alternative Exchange Rate Regimes", *Journal of Monetary Economics*, XXIII, pp. 401-431

Grossman, G. and Helpman, E. [1989]: “Product Development and International Trade”, *Journal of Political Economy* XCVII, pp. 1261-1283

Grossman, G.M. and Helpman E. [1990]: “Comparative Advantage and Long-Run Growth,” *American Economic Review* LXXX, pp. 796-815

Grossman, G.M. and Helpman E. [1991]: “Quality Ladders and Product Cycles,” *Quarterly Journal of Economics* CVI, pp. 557-586

Grossman, G. and Helpman, E. [1991a]: “Quality Ladders in the Theory of Growth”, *Review of Economic Studies* LVIII, pp. 43-61

Grossman, G. M. and Helpman, E.: “Innovation and Growth in the Global Economy”, MIT Press, Cambridge MA, 1991 ISBN 0-262-07136-3

Hall, R. E. and Jones, C. I. [1999]: “Why Do Some Countries Produces More Output per Worker than Others?” *Quarterly Journal of Economics*

Heathcote, J. and Perri, F. [2002]: “Financial Autarky and International Business Cycles”, *Journal of Monetary Economics*, XLIX, pp. 601-627

Helpman, E. [1993]: “Innovation, Imitation, and Intellectual Property Rights,” *Econometrica* LXI, pp. 1247-1280

Helpman, E. and Krugman, P. R.: *Market Structure and Foreign Trade*, MIT Press, Cambridge, Massachusetts; London, England 1985 ISBN 0-262-08150-4

Howitt, P. [2000]: “Endogenous Growth and Cross-Country Income Differences,” *American Economic Review* CXC, pp. 829-846

Jensen, R. and Thursby, M. [1987]: “A Decision Theoretic Model of Innovation, Technology Transfer and Trade”, *Review of Economic Studies*, LIV, pp.631-647

Klenow, P. J. and Rodriguez-Clare A. [1997]: "The Neoclassical Revival in Growth Economics: Has It Gone Too Far?" NBER Macroeconomics Annual, MIT Press, Cambridge, MA, pp. 73-102

Klundert, T. van de and Smulders, S. [1996]: “North-South Knowledge Spillovers and Competition: Convergence versus Divergence,” *Journal of Development Economics* L, pp. 213-232

Klundert, T. van de and Smulders, S. [2001]: “Loss of Technological Leadership of Rentier Economies: a Two-Country Endogenous Growth Model,” *Journal of International Economics* LIV, pp. 211-231

Kraay, A., Loayza, N., Servén L. and Ventura J. [2005]: “Country Portfolios”, *Journal of the European Economic Association*, III, forthcoming

Krugman, P. [1979]: “A Model of Innovation, Technological Transfer and the World Distribution of Income”, *Journal of Political Economy*, LXXXVII, pp. 253-266

Krugman, P. [1981]: “Trade, Accumulation, and Uneven Growth,” *Journal of International Economics* VIII, pp. 149-161

Krugman, P. [1987]; “The Narrow Moving Band, the Dutch Disease, and the Competitive Consequences of Mrs. Thatcher,” *Journal of Development Economics* XXVII, pp. 41-55

Krugman, P. [1991] “Increasing Returns and Economic Geography”, *Journal of Political Economy*, IC, pp. 483-499

Krugman, P. [1991]: “History vs. Expectations”, *Quarterly Journal of Economics*, CVI, pp. 651-667.

Krugman, P., and A. Venables [1995]: “Globalization and the Inequality of Nations,” *Quarterly Journal of Economics*, CIX, pp. 857-880.

Lai, E.L.-C. [1995]: “The Product cycle and the World Distribution of Income. A Reformulation,” *Journal of International Economics* IXL, pp. 369-382

Lai, E. L.-C. [1998]; “International Intellectual Property Rights Protection and the Rate of Product Innovation” *Journal of Development Economics* LV, pp. 133-153

Levchenko, A. [2004]: “Institutional Quality and International Trade”, manuscript MIT

Loayza N. V., Knight M., and Villanueva D. [1993]: “Testing the Neoclassical Theory of Economics Growth: A Panel Data Approach” *IMF Staff Papers*

Lundborg, P. and Segerstrom, P.S. [2002]: “The Growth and Welfare Effects of International Mass Migration,” *Journal of International Economics* LVI, pp. 177-204

Mankiw, N. G., Romer, D., and Weil, D. N. [1992]: “A Contribution to the Empirics of Economic Growth” *Quarterly Journal of Economics* Vol. 107, No. 2, pp. 407-437

Martin, P. and Rey H. [2002]: “Globalization and Emerging Markets: With or Without Crash”, CEPR working paper DP3378

Martin, P. and Rey H. [2004]: “Financial Super-Markets: Size Matters for Asset Trade,” *Journal of* In*ternational Economics*, LXIV, pp. 335-661.

Matsuyama, K. [1991]: “Increasing Returns, Industrialization and Indeterminacy of Equilibrium,” *Quarterly Journal of Economic*s, CVI, pp. 617-650.

Matsuyama, K. [2000]: “A Ricardian Model with a Continuum of Goods under Nonhomothetic Preferences: Demand Complementarities, Income Distribution, and North-South Trade”, *Journal of Political Economy*, CVIII, pp. 1093-1120

Matsuyama, K. [2004]: “Financial Market Globalization, Symmetry-Breaking and Endogenous Inequality of Nations”, *Econometrica*, LXXII, No. 3, pp.853-884

Matsuyama, K. [2004a]: “Beyond Icebergs: Modeling Globalization as Biased Technical Change” *Working Paper* Northwestern University

Modigliani, F. and Ando, A. [1963]: "The Life Cycle Hypothesis of Saving", *American Economic Review* Vol. 53, No. 1, pp. 55-84

Molana, H., Vines, D. [1989]: “North-South Growth and the Terms of Trade: A Model on Kaldorian Lines”, *Economic Journal*, IC pp. 443-453

Mountford, A. [1998]: “Trade, Convergence and Overtaking,” *Journal of International Economics* XLVI, pp. 167-182.

Myers, M. G. [1970]: ”Equilibrium Growth and Capital Movements Between Open Economies,” *American Economic Review* LX, pp. 393-397.

Dixit, A.K., and V. Norman: *Theory of International Trade*, Cambridge University Press, Cambridge, New York, Port Chester, Melbourne, Sydney 1989, ISBN 0521234816

Obstfeld, M. [1989]: “Fiscal Deficits and Relative Prices in a Growing World Economy”, *Journal of Monetary Economics*, XXIII, pp. 461-484

Obstfeld, M. [1994]: “Risk-Taking, Global Diversification, and Growth,” *American Economic Review* LXXXIV, pp. 1310-1329.

Oniki, H. and Uzawa H. [1965]: “Patterns of Trade and Investment in a Dynamic Model of International Trade” *Review of Economics Studies* XXXII, pp. 15-38.

Ono, Y. and Shibata, A. [1991]: “Spill-Over Effects of Supply-Side Changes in a Two-Country Economy with Capital Accumulation,” *Journal of International Economics* XXXIII, pp. 127-146.

Ortega, F. [2004]: “Immigration policy and the Welfare State”, manuscript UPF

Pritchett, L. [1997]: “Divergence, Big Time”, *Journal of Economic Perspectives,* Vol. 11, No. 3, pp. 3-17

Puga, D., Venables, A. J. [1999]: “Agglomeration and Economic Development: Import Substitution vs. Trade Liberalisation”, *Economic Journal*, CIX, pp. 292-311

Rauch, J. E. [1991]: “Reconciling the Pattern of trade with the Pattern of Migration,” *American Economic Review* LXXXI, pp. 775-796

Rivera-Batiz, L. and Romer P. A. [1991]: “Economic Integration and endogenous Growth,” *Quarterly Journal of Economics* CVI, pp. 531-555.

Rodriguez, F. and Rodrik, D. [2000]: “Trade Policy and Economic Growth: A Skeptic’s Guide to the Cross-National Evidence” *NBER Macroeconomic Annual*

Ruffin, R.J. [1979]: “Growth and the Long-Run Theory of International Capital Movements,” *American Economic Review* LXIX, pp. 832-842

Sachs; J. and Warner, A. [1995]: "Economic Reform and the Process of Global Integration" Brookings Papers on Economic Activity, No.1

Samuelson, P. A. [1948]: “International Trade and The Equalization of Factor Prices”, *Economic Journal*, LVIII, pp. 163-84

Samuelson, P. A. [1949]: “International Trade and The Equalization of Factor Prices, Again”, *Economic Journal*, Vol. 59, No. 234, pp. 181-197

Sauré, P. [2004]: “Revisiting the Infant Industry Argument”, manuscript UPF

Sauré, P. [2004]: “How to Use Industrial Policy to Sustain Trade Agreements”, manuscript UPF

Segerstrom, P. S., Anant, T.C.A., and E. Dinopoulos [1990]: “A Schumpeterian Model of the Product Life Cycle,” *American Economic Review* LXXX, pp. 1077-1091.

Şener, F. [2001]: “Schumpeterian Unemployment, Trade and Wages,” *Journal of International Economics* LIV, pp. 119-148.

Sibert, A. [1985]: “Capital Accumulation and Foreign Investment Taxation”, *Review of Economic Studies*, LII, pp. 331-345

Solow, R. M. [1956]: “A Contribution to the Theory of Economic Growth”, *Quarterly Journal of Economics*, Vol. 70, No. 1, pp. 65-94

Stiglitz, J. E. [1970]: “Factor Price Equalization in a Dynamic Economy,” *Journal of Political Economy* LXXVIII, pp. 456-488

Stokey, N. L. [1991]: “The Volume and Composition of Trade Between Rich and Poor Countries”, *Review of Economic Studies*, LVIII, pp. 63-80

Taylor, M. S. [1993]: “Quality Ladders and Ricardian Trade,” *Journal of International Economics* XXXIV, pp. 225-243.

Taylor, M. S. [1994]: “”Once-Off and Continuing Gains from Trade”, *Review of Economic Studies*, LXI, pp. 589-601

Trefler, D. [1993]: "International Factor Price Differences: Leontief was Right!" *Journal of Political Economy* Vol. 101, No. 6, pp. 961-987

Vanek, J. [1971]: “Economic Growth and International Trade in Pure Theory,” *Quarterly Journal of Economics* LXXXV, pp. 377-390.

Ventura, J. [1997], “Growth and Interdependence,” *Quarterly Journal of Economics*, CXII, 1, pp. 57-84.

Ventura, J. [2002], “Bubbles and Capital Flows” NBER Working Paper No. 9304

Wang, J.-Y. [1990]: “Growth, Technology transfer, and the Long-Run Theory of International Capital Movements,” *Journal of International Economics* XXIX, pp. 255-271.

Wilson, C. [1980]: “On the General Structure of Ricardian Models with a Continuum of Goods: Applications to Growth, Tariff Theory, and Technical Change” *Econometrica*. 48: pp. 1675-1702

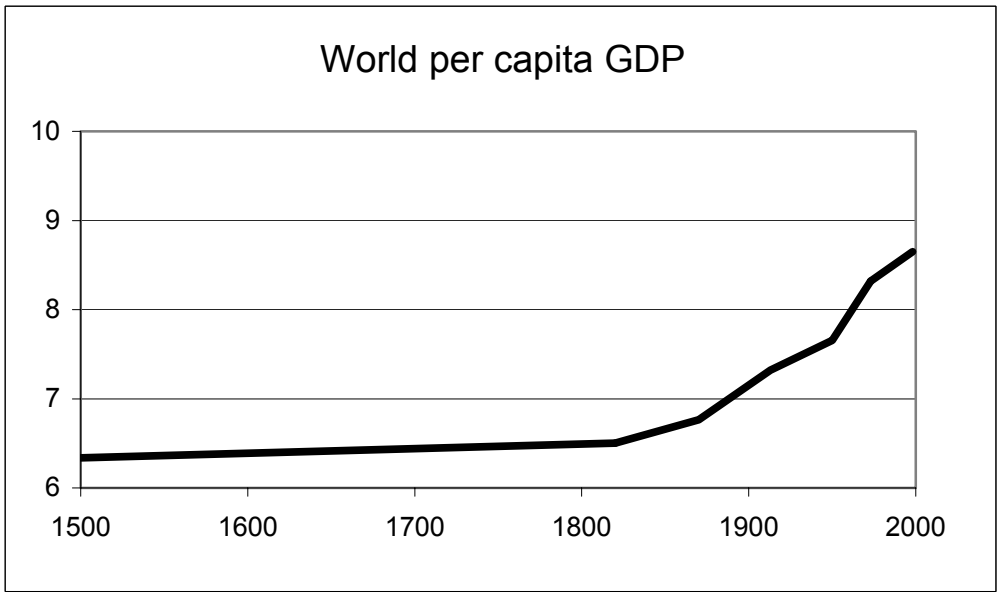
Yanagawa, N. [1996]: “Economic Development in a World with Many Countries,” *Journal of Development Economics* IL, pp. 271-288.

Yang, G. and Maskus K.E. [2001]: “Intellectual Property Rights, Licensing, and Innovation in an Endogenous Product-Cycle Model,” *Journal of International Economics* LIII, pp. 169-387.

Young, A. [1991]: “Learning by Doing and the Dynamic Effects of International Trade” *Quarterly Journal of Economics*, Vol. 106, No. 2, pp. 369-405

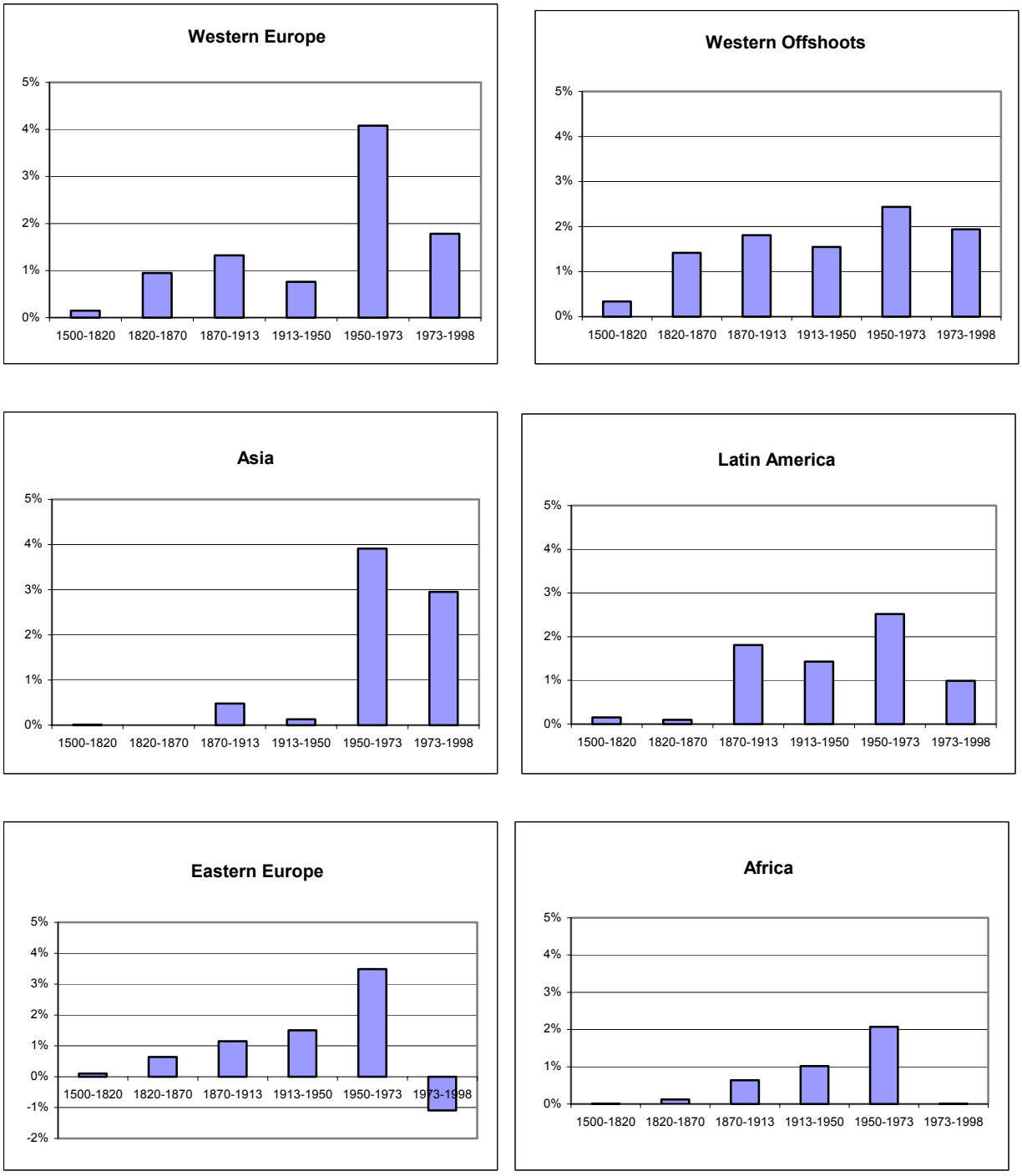
Young, A. [1995]: “The Tyranny of Numbers: Confronting the Statistical Realities of the East Asian Growth Experience” *Quarterly Journal of Economics*, Vol. 110, No. 3, pp. 641-680

Figure 1



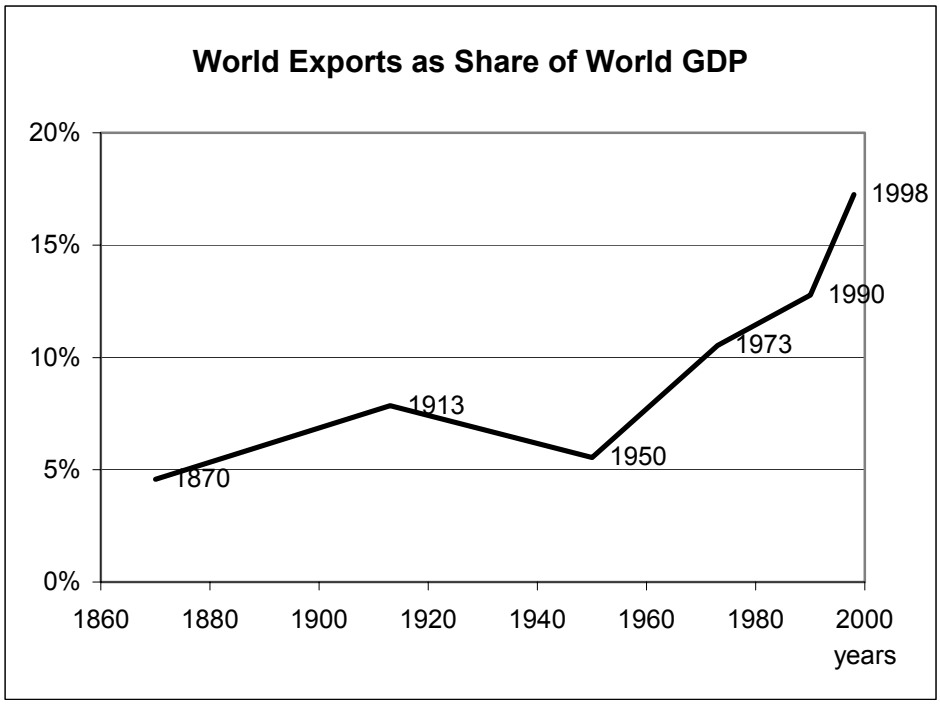
Notes: This figure shows the dynamics of world per capita GDP for the selected years 1500, 1820, 1870, 1913, 1950, 1973, and 1998 (in log of 1990 US$). Data are from Angus Maddison, "The World Economy – A Millennial Perspective" Table 3-1b page126.

Figure 2
Per capita GDP Growth



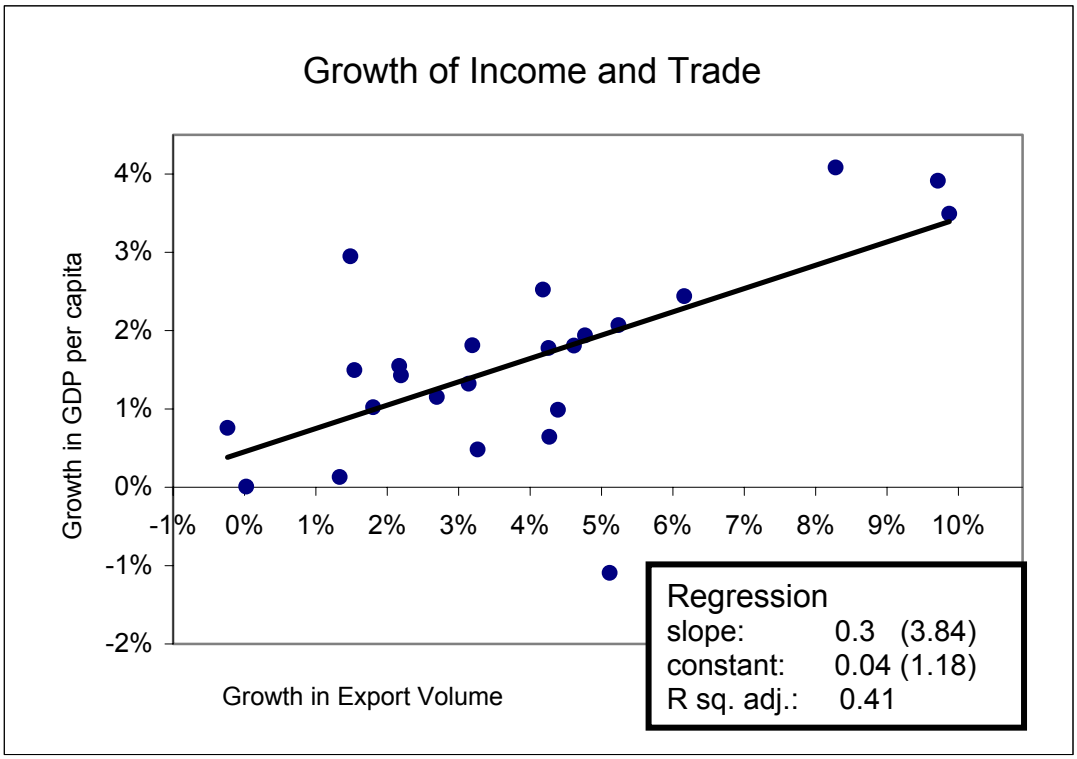
Notes: This figure shows average annual growth rates by major world regions for selected periods. Data are derived from Angus Maddison, "The World Economy – A Millennial Perspective" Table 3-1b page126. (Western Europe contains Austria, Belgium, Denmark, Finland, France, Germany, Italy, Netherlands, Norway, Sweden, Switzerland, UK, Portugal, Spain, Greece and 13 small countries; Western Offshoots are United States, Canada, Australia and New Zealand; Asia is China, India, Japan, Korea, Indonesia, Indochina, Iran, Turkey and Other East and West Asian countries; Latin America includes Brazil, Mexico, Peru, and Others; Eastern Europe contains Albania, Bulgaria, Hungary, Poland, Romania and territories of former Czechoslovakia and Yugoslavia; Africa is Egypt and Others.)

Figure 3



Notes: The figure shows Volume of World Exports over World GDP (in constant US$) for selected dates. Data are from Tables 3-1b, A1-b, A2-b, A3-b, A4-b, pages126, 184, 194, 214, and 223 in Angus Maddison, "The World Economy – A Millennial Perspective".

Figure 4

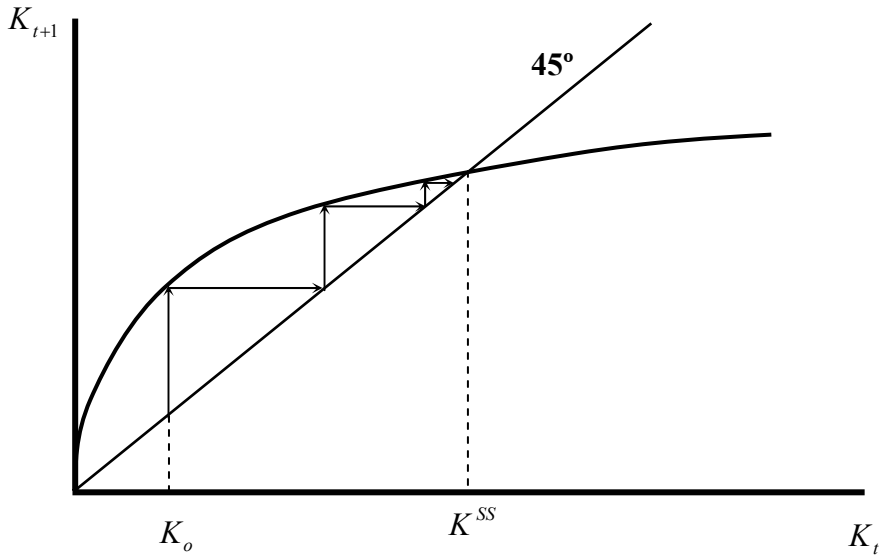


Notes: This figure plots annualized rate of trade growth against annualized rate of per capita GDP growth for major world regions and selected periods. The Regions are Western Europe, Western Offshoots, Eastern Europe and former USSR, Latin America, Asia, and Africa. Periods are 1870-1913, 1913-1950, 1950-1973 and 1973-1998. Each data point stands for one region during one period. The solid line represents the prediction of a linear regression. The estimated regression are reported in the box, t-statistics are in brackets. Data are from Angus Maddison, "The World Economy – A Millennial Perspective". Data for GDP growth are obtained from Table 3-1b page126 and Table B-10 page 241 (to include Japan). Data for export growth are derived from Table F-3 page362 and Tables A1-b, A2-b, A3-b, and A4-b, pages 184, 194, 214 and 223, respectively.

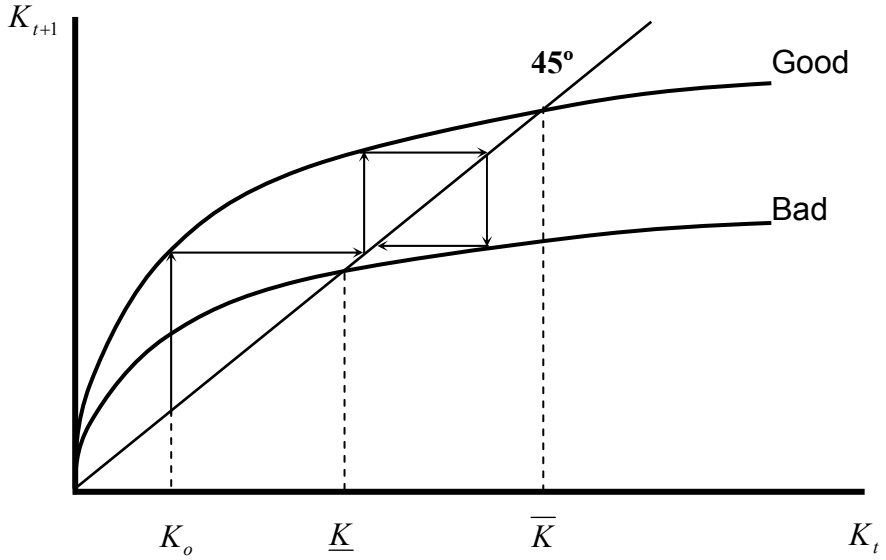
Figure 5

$\alpha\mu + \upsilon < 1$

The “deterministic” case



The “stochastic” case

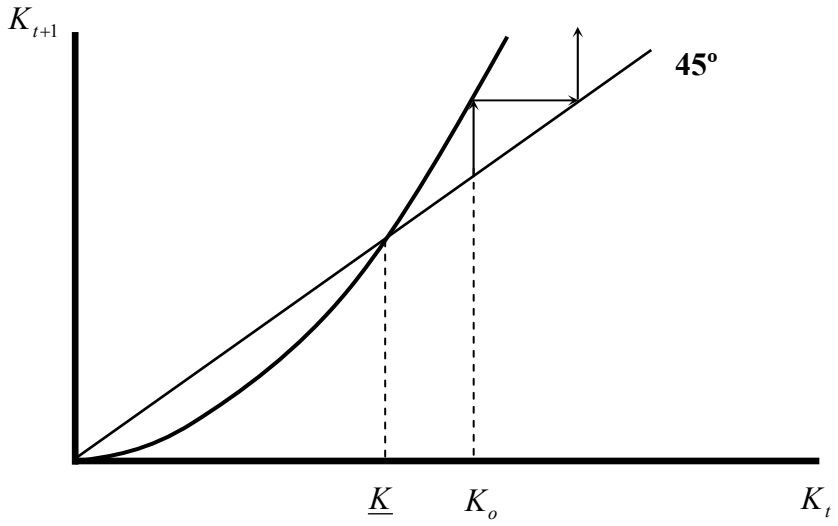


Notes: This figure shows the case of strong deminishing returns and weak market size effects. In the top panel, the stock of physical capital converges monotonically to its unique steady state. The bottom panel shows the stochastic case, where the stock of physical capital converges to the steady state interval $\left[\underline{K}, \overline{K}\right]$ within which it fluctuates according to the states of the world.

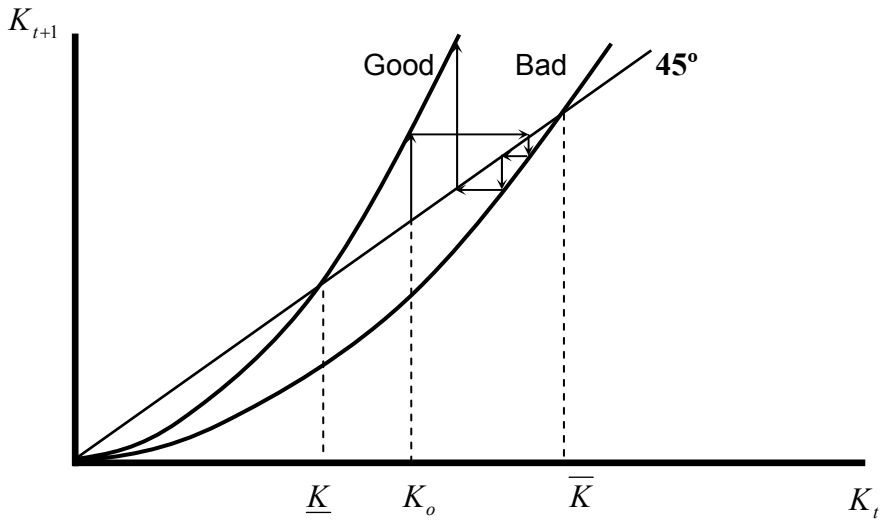
Figure 6

$$\alpha\mu + \upsilon > 1$$

The "deterministic" case



The "stochastic" case

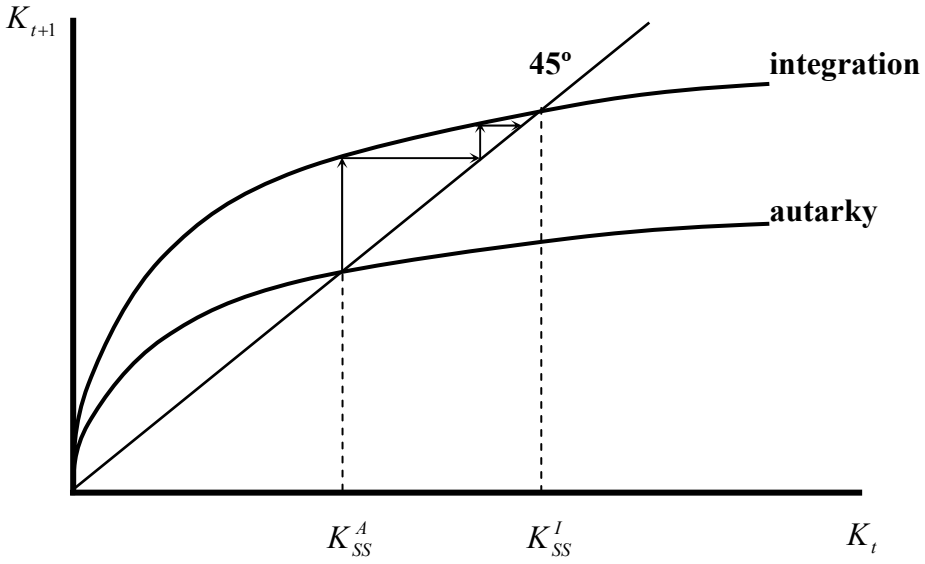


Notes: This figure shows the case of weak diminishing returns and strong market size effects. In the top panel, the stock of physical capital grows at increasing rates since $K_o > \underline{K}$. In the bottom panel the stock of physical capital fluctuates between $\underline{K}$ and $\overline{K}$ according to the states of the world, until it eventually leaves this range.

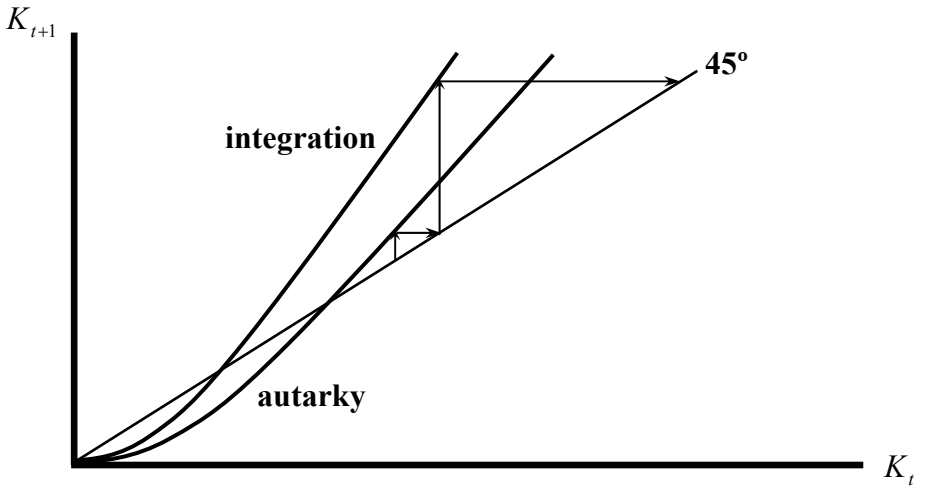
Figure 7

Effects of Economic Integration

$\alpha\mu + \upsilon < 1$

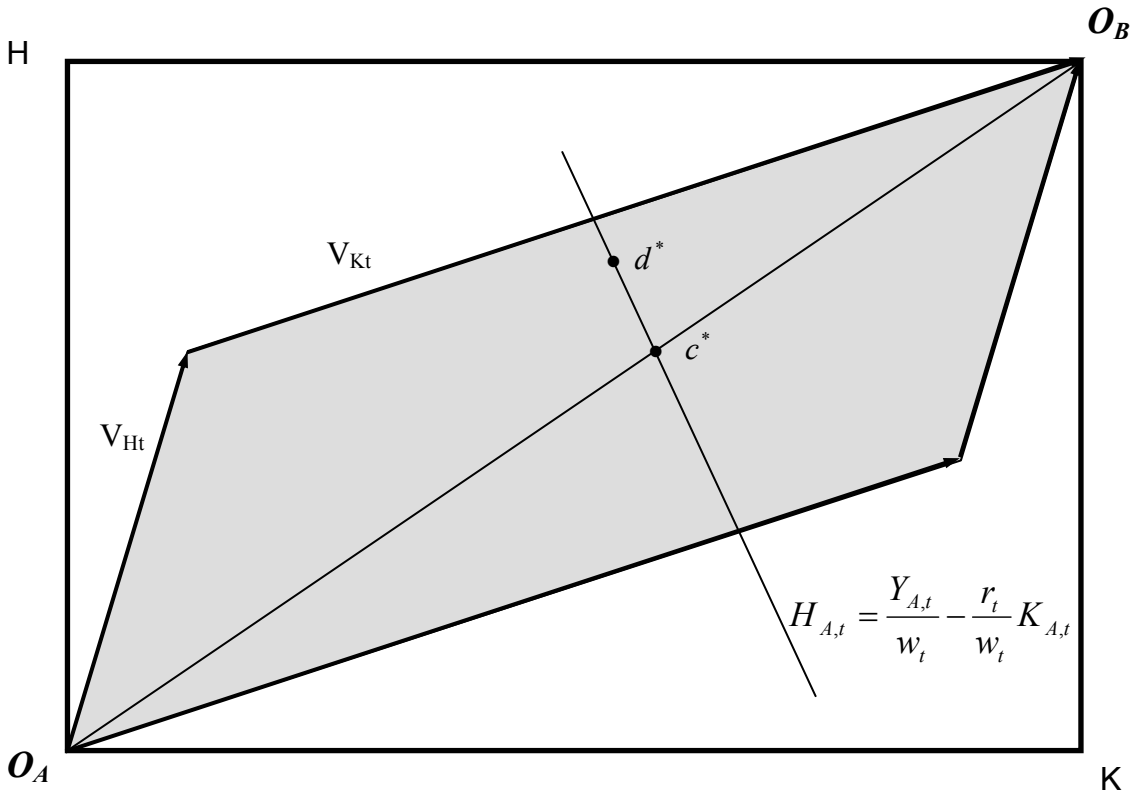


$\alpha\mu + \upsilon > 1$



Notes: This figure illustrates the effects of economic integration. The top panel shows that, if $\alpha \cdot \mu + \upsilon < 1$, economic integration has level effects on income. The bottom panel shows that, if $\alpha \cdot \mu + \upsilon > 1$, economic integration has growth effects on incomes.

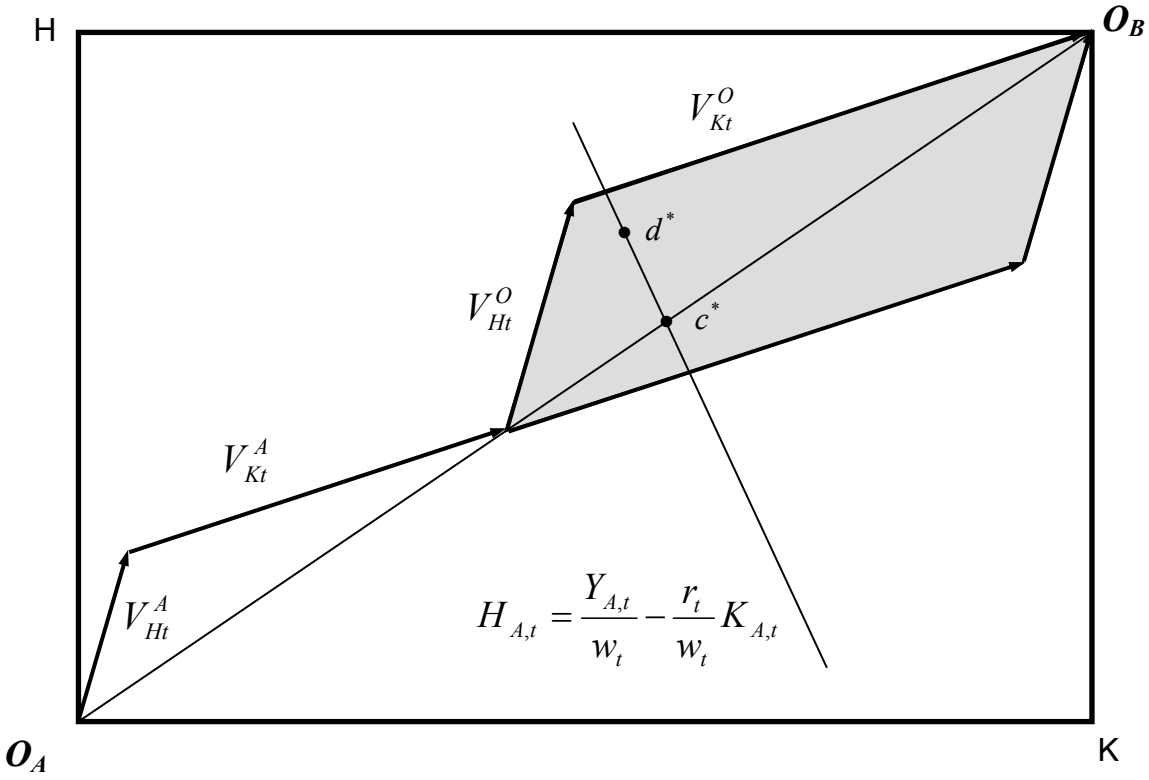
Figure 8



Notes: The box in this figure is a geometrical representation of the set D_t, as each element of this set is a point in the box and vice versa. For instance, d^* is a factor distribution such that A-regions have more human and physical capital than B-regions; but human capital is relatively more abundant in A-regions than in B-regions. The box also contains a set of vectors that represent the factor usage per industry that would apply in the integrated economy. For instance, the vector V_{it} has height H_{it} and width K_{it}. The set FPE_t is the gray area. Since all regions have the same industry productivities, production trivially takes place only in regions with the highest possible productivity (requirement R1). Each of the points in the gray area can be generated as a convex combination of the integrated economy's vectors of factor usage per industry (requirement R2). Since β_i=0, trivially there are no fixed costs of production that are incurred twice (requirement R3). Points outside of the shaded area do not have this property and therefore do not belong to FPE_t.

The factor content of production is given by the regions' factor endowments, i.e. d^*. Since all regions have the same spending shares and use the same techniques to produce all goods, the factor content of consumption lies in the diagonal, i.e. c^*.
In A-regions, the H-industry is a net exporter while the K-industry is a net importer. The opposite occurs in B-regions.

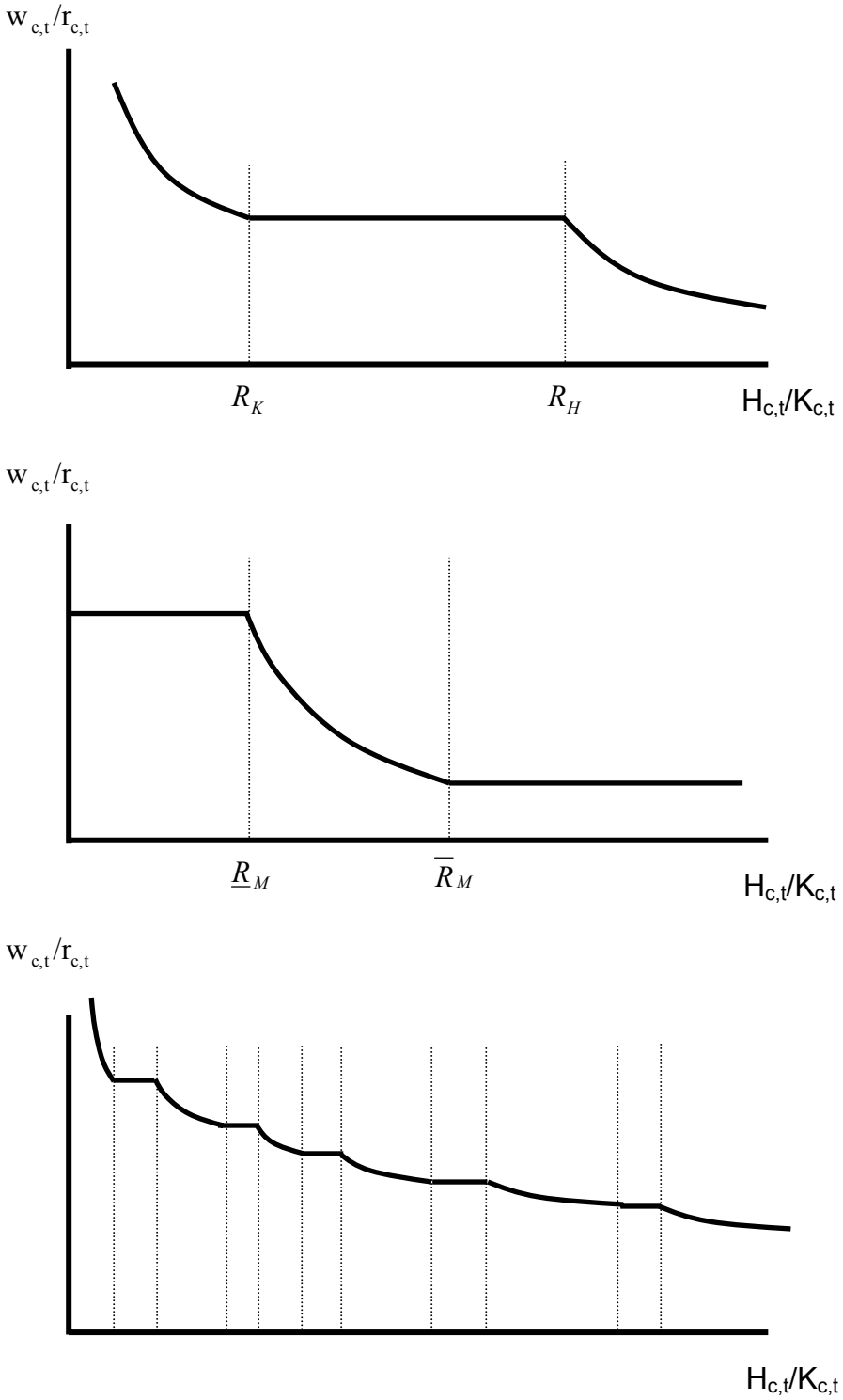
Figure 9



Notes: The box in this Figure is a geometrical representation of the set D_t, as each element of this set is a point in the box and vice versa. For instance, d^* is a factor distribution such that A-regions have more human and physical capital than B-regions; but human capital is relatively more abundant in A-regions than in B-regions. There are four different industries, "advanced" physical (human) capital intensive and "backward" physical (human) capital intensive. The A-countries have a highest productivity in the "advanced" industries; technologies in the "backward" industries are equal in all countries. The vectors V^X_{it} have height H^X_{it} and width K^X_{it} and represent the factor content of the *X*-industries, where *X=A,B* stands for "advanced" or "backward" industries. The set FPE_t is the shaded area. In this set, all "advanced" industries must be located in the A-countries (requirement R1). Once this requirement is satisfied, each of the points in the shaded area can be generated as a convex combination of the integrated economy's vectors of factor usage of the "backward" industries (requirement R2). Since β_i=0, trivially there are no fixed costs of production that are incurred twice (requirement R3). Points outside of the shaded area do not have both properties and therefore do not belong to FPE_t.

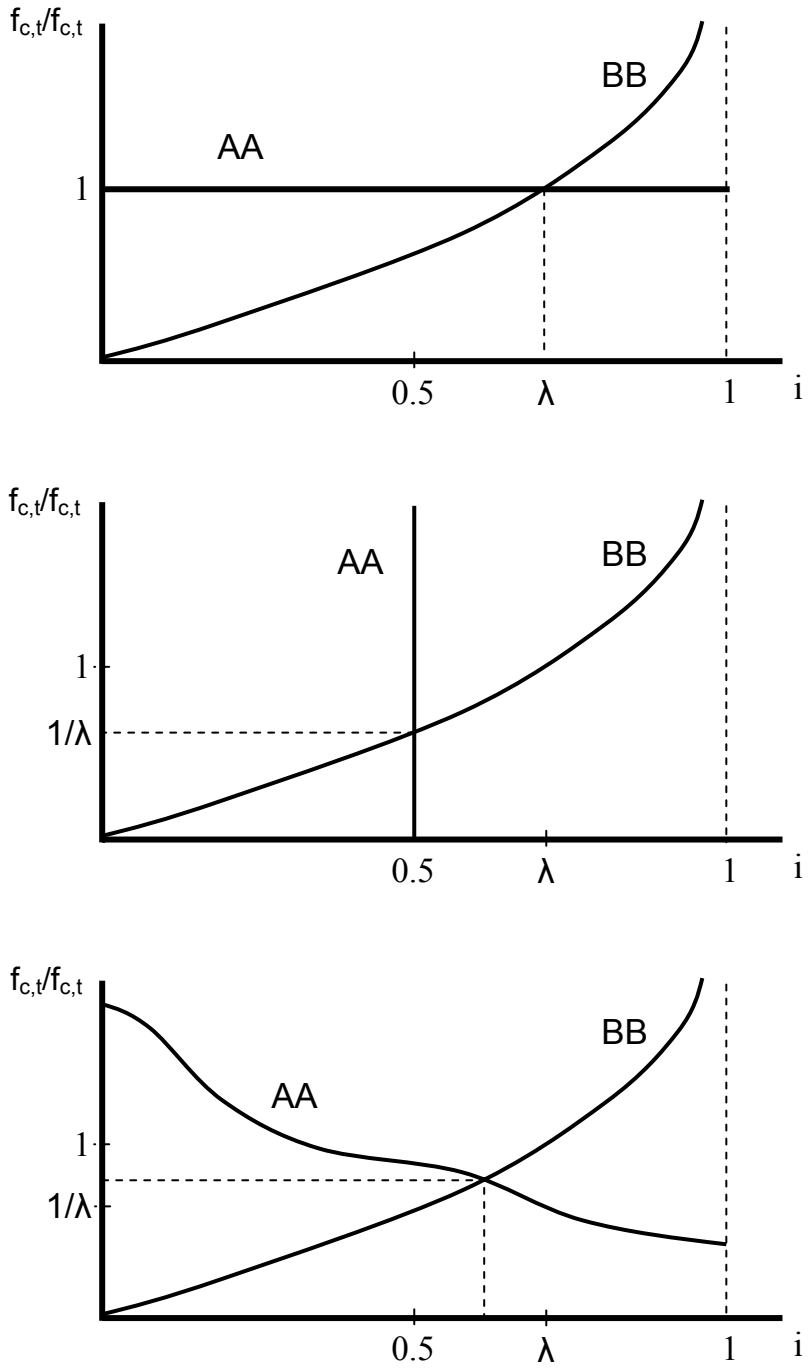
The factor content of production is given by the regions' factor endowments, i.e. d^*. Since all regions have the same spending shares and use the same techniques to produce all goods, the factor content of consumption lies in the diagonal, i.e. c^*. In H-regions, the H-industry is a net exporter while the K-industry is a net importer. The opposite occurs in K-regions.

Figure 10



Notes: This figure shows how the wage-rental ratio varies with the factor proportions. The top panel represents a two-goods, one-cone world where countries with extreme factor proportions are outside the cone (Example 2.4.1). The middle panel represents a three-good, two-cone world where countries with intermediate factor proportions lie outside the cone (Example 2.4.2). The bottom panel shows a world with multiple goods and cones.

Figure 11



Notes: This figure shows how pattern of production and trade (i*) and relative factor costs ($f_{N,t}/f_{S,t}$) are determined in Example 2.5.3. The top panel shows the case of arbitrarily small differences in industry productivities. The middle panel shows the case of arbitrarily large differences in industry productivities. The bottom panel shows the intermediate case.

Trade, Growth and the Size of Countries*

Alberto Alesina
Harvard University
CEPR and NBER

Enrico Spolaore
Brown University

Romain Wacziarg
Stanford University
and NBER

August 2004
Forthcoming, Handbook of Economic Growth

Abstract

Normally, economists take the size of countries as an exogenous variable. Nevertheless, the borders of countries and therefore their size change, partially in response to economic factors such as the pattern of international trade. Conversely, the size of countries influences their economic performance and their preferences for international economic policies - for instance smaller countries have a greater stake in maintaining free trade. In this paper, we review the theory and evidence concerning a growing body of research that considers both the impact of market size on growth and the endogenous determination of country size. We argue that our understanding of economic performance and of the history of international economic integration can be greatly improved by bringing the issue of country size at the forefront of the analysis of growth.

*Alberto Alesina: Department of Economics, Harvard University, Cambridge, MA 02138, aalesina@kuznets.fas.harvard.edu. Enrico Spolaore: Department of Economics, Brown University, Providence, RI 02912, enrico_spolaore@brown.edu. Romain Wacziarg: Stanford Graduate School of Business, 518 Memorial Way, Stanford CA 94305, wacziarg@gsb.stanford.edu. This paper was prepared for the Handbook of Economic Growth, edited by Philippe Aghion and Steven Durlauf, North Holland. We are grateful to the NSF for financial support with a grant through the NBER. We also thank Jessica Seddon Wallack for excellent research assistance. We thank Philippe Aghion, Francisco Alcalá, Michele Boldrin, Antonio Ciccone, and seminar and conference participants at the University of Modena, Harvard University, and the European University Institute for useful comments. All remaining errors are ours.

1 Introduction

Does size matter for economic success? Of the five largest countries in the world in terms of population, China, India, the United States, Indonesia and Brazil, only the United States is a rich country.[1] In fact the richest country in the world in 2000, in terms of income per capita, was Luxembourg, with less than 500,000 inhabitants. Among the richest countries in the world, many have populations well below the world median, which was about 6 million people in 2000. And when we consider growth of income per capita rather than income levels, again we find small countries among the top performers. For example Singapore, with 3 million inhabitants, experienced the highest growth rate of per capita income of any country between 1960 and 1990.[2] These examples show that a country can be small and prosper, or, at the very least, that size alone is not enough to guarantee economic success.

In this paper, we discuss the relationship between the scale of an economy and economic growth from two points of view. We first discuss the effects of an economy's size on its growth rate and we then examine how the size of countries evolves in response to economic factors.

The "new growth literature", with its emphasis on increasing returns to scale, has devoted much attention to the question of size of an economy.[3] It is therefore somewhat surprising that the question of the effect of border design and size of the polity as a determinant of economic growth has received limited attention. One reason is that, as we will see below, measures of country size (population or land area) used alone in growth regressions, generally do not have much explanatory power. Even less attention has been devoted to the endogenous determination of borders even from those researchers who have paid attention to the effect of geography on growth. Borders are not exogenous geographical features: they are a man-made institution. In fact, even the geographical characteristics of a country are in some sense endogenous: for instance whether a country is landlocked or not

[1] Throughout this paper we use the word "country", "nation" and "state" interchangeably, meaning a polity defined by borders and a national government and citizens. We are not dealing with the concept of a nation as a people not necessarily identified by borders and a government.

[2] Based on all measures of growth in per capita PPP income in constant prices constructed from the Penn World Tables version 6.1.

[3] However, it is well known that increasing returns are not *necessary* for a positive relationship between market size and economic performance. As we will see in our analytical section, larger markets may entail larger gains from trade and higher income per capita even when the technology exhibits constant returns to scale.

is the result of the design of its borders, which in turn depend upon domestic and international factors.

While economists have remain on the sidelines on this topic, philosophers devoted much energy thinking about country size. Plato, Aristotle and Montesquieu worried that a large polity cannot be run as a democracy. Aristotle wrote in Politics that "experience has shown that it is difficult, if not impossible, for a populous state to be run by good laws". Influenced by Montesquieu, the founding fathers of the United States were preoccupied with the potentially excessive size of the new Federal State. On the other hand, liberal thinkers who in the nineteenth century contributed to defining modern nation-states were concerned that in order to be economically, and therefore politically viable, countries should not be too small. Historians have studied the formation of states and their size and emphasized the role of wars and military technology as an important determinant. In fact, rulers, especially non-democratic ones, have always seen size as a measure of power and tried to expand the size of the territory under their rule. So, while throughout history country size seemed to be a constant preoccupation of philosophers, political scientists and policymakers, economists have largely ignored this subject.

In recent decades the question of borders has risen to the center of attention in international politics. The collapse of the Soviet Union, decolonization, and the break-up of several countries have rapidly increased the number of independent polities. In 1946 there were 76 independent countries, in 2002 there were 193.[4] East Timor was the latest new independent country at the time of this writing.

In this paper, we explore the relatively small recent economics literature dealing with the size of countries and its effect on economic growth. In particular we ask several questions: does size matter for economic success, and if so why and through which channels? What forces lead to changes in the organization of borders, or to put it differently what determines the evolution of the size of countries? Obviously the second question is very broad. Here we focus specifically a narrower version of this question, namely how economic factors, especially the trade regime, influence size.[5]

This paper is organized as follows. Section 2 discusses a general framework for thinking in economic terms about the optimal and the equilibrium

[4] These include the 191 member states of the United Nations, plus the Vatican and Taiwan.

[5] For a broader discussion see Alesina and Spolaore (2003).

size of countries, providing a formal model that focuses on the effect of size on income levels and growth, with special emphasis on the role of trade. Section 3 reviews the empirical evidence on these issues and provides updated and new results. Section 4 briefly explores how the relationship between country size, international trade and growth have played out historically. The last section highlights questions for future research.

2 Size, Openness and Growth: Theory

2.1 The costs and benefits of size

We think of the equilibrium size of countries as emerging from the trade-off between the benefit of size and the costs of preference heterogeneity in the population, an approach followed by Alesina and Spolaore (1997, 2003) and Alesina, Spolaore and Wacziarg (2000).

2.1.1 The Benefits of Size

The main benefits from size in terms of population are the following:

1) There are economies of scale in the production of public goods. The per capita cost of many public goods is lower in larger countries, where more taxpayers pay for them. Think, for instance, of defense, a monetary and financial system, a judicial system, infrastructure for communications, police and crime prevention, public health, embassies, national parks, etc. In many cases, part of the cost of public goods is independent of the number of users or taxpayers, or grows less than proportionally, so that the per capita costs of many public goods is declining with the number of taxpayers. Alesina and Wacziarg (1998) documented that the share of government spending over GDP is decreasing in population; that is, smaller countries have larger governments.

2) A larger country (both in terms of population and national product) is less subject to foreign aggression. Thus, safety is a public good that increases with country size. Also, and related to the size of government argument above, smaller countries may have to spend proportionally more for defense than larger countries given economies of scale in defense spending. Empirically, the relationship between country size and share of spending of defense is affected by the fact that small countries can enter into military alliances, but in general, size brings about more safety. Note that if a small country enters into a military alliance with a larger one, the latter may

provide defense, but it may extract some form of compensation, direct or indirect, from the smaller partner. In this sense, even allowing for military alliances, being large is an advantage.

3) Larger countries can better internalize cross-regional externalities by centralizing the provision of those public goods that involve strong externalities.[6]

4) Larger countries are better able to provide insurance to regions affected by imperfectly correlated shocks. Consider Catalonia, for instance. If this region experiences a recession worse than the Spanish average, it receives fiscal and other transfers, on net, from the rest of the country. Obviously, the reverse holds as well. When Catalonia does better than average, it becomes a net provider of transfers to other Spanish regions. If Catalonia, instead, were independent, it would have a more pronounced business cycle because it would not receive help during especially bad recessions, and would not have to provide for others in case of exceptional booms.[7]

5) Larger countries can build redistributive schemes from richer to poorer regions, therefore achieving distributions of after tax income which would not be available to individual regions acting independently. This is why poorer than average regions would want to form larger countries inclusive of richer regions, while the latter may prefer independence.[8]

6). Finally, the role of market size is the issue on which we focus most in this article. Adam Smith (1776) already had the intuition that the extent of the market creates a limit on specialization. More recently, a well established literature from Romer (1986), Lucas (1988) to Grossman and Helpman (1991) has emphasized the benefits of scale in light of positive externalities in the accumulation of human capital and the transmission of knowledge, or in light of increasing returns to scale embedded in technology or knowledge creation.[9] Murphy, Shleifer and Vishny (1987) focused instead

[6] See Alesina and Wacziarg (1999) for a discussion of this point in the context of Europe. For example, fisheries policy has been centralized in Europe because if each country decided on its own fishing policy, the result would be overfishing and resource depletion. For some policies, such as policies to limit global warming, centralization at the world level might be justified.

[7] Obviously, this argument relies on an assumption that international capital markets are imperfect, so that independenct countries cannot fully self-insure..

[8] See Bolton and Roland (1997) for a theoretical treatment of this point.

[9] A recent critique of some this class of models is due to Jones (1995b). Specifically, Jones pointed out that endogenous growth models generally imply that growth rates should increase with the stock of knowledge. Yet growth rates have been relatively stable or

on the benefits of size in models of "take-off" or "big push" of industrialization, where the take-off phase is characterized by a transition from a slow growth, constant returns to scale technology to an endogenous growth, increasing returns to scale technology. Finally, several papers have stressed the pro-competitive effects of a larger market size: size enhances growth by raising the intensity of product market competition.[10] In these various models, size represents the stock of individuals, purchasing power and income that interact in the market. This market may or may not coincide with the political size of a country as defined by its borders. It does coincide with it if a country is completely autarkic, i.e. does not engage in exchanges of goods or factors of production with the rest of the world. On the contrary, market size and country size are uncorrelated in a world of complete free trade. So in models with increasing returns to scale, market size depends both on country size and on trade openness.

In theory, with no obstacle to the cross-border circulation of factors of productions, goods and ideas, country size should be, at least through the channel of market size, irrelevant for economic success. Thus, in a world of free trade, redrawing borders should have no effect on economic efficiency and productivity. However, a vast literature has convincingly shown that even in the absence of explicit trade policy barriers, crossing borders is indeed costly, so that economic interactions within a country are much easier and denser than across borders. This is true both for trade in goods and financial assets.[11] What explains this border effect, even in the absence of explicit policy barriers, is not completely clear.[12] Whatever the source of the border effect, however, the correlation between the "political size" of a country and its market size does not totally disappear even in the absence of policy-induced trade barriers. Still, one would expect that the correlation between size and economic success is mediated by the trade regime. In a regime of free trade, small countries can prosper, while in a world of trade barriers, being large is much more important for economic prosperity, measured for instance by income per capita.

declining in advanced industrial economies, while the stock of knowledge has increased rapidly. In section 3, we review and discuss this critique in much detail.

[10] See Aghion and Howitt (1998) and Aghion et al. (2002).

[11] On trade see McCallum (1985), Helliwell (1998). For the role of geographical factors in financial flows, see Portes and Rey (1999). For a theoretical discussion of transportation costs across borders and their effects on market integration, see Obstfeld and Rogoff (2000).

[12] A recent literature prompted by Rose (2000) argues that not having the same currency creates large trade barriers. For a review of the evidence see Alesina, Barro and Tenreyro (2002). Other explanatory factors include different languages, different legal standards, difficulties in enforcing contracts across political borders, etc.

2.1.2 The Costs of Size

If size only had benefits, then the world should be organized as a single political entity. This is not the case. Why? As countries become larger and larger, administrative and congestion costs may overcome the benefits of size pointed out above. However, these types of costs become binding only for very large countries and they are not likely to be relevant determinants of the existing countries, many of which are quite small. As we noted above, the median country size is less than six million inhabitants.

A much more important constraint on the feasible size of countries lies in the heterogeneity of individuals' preferences. Being part of the same country implies sharing public goods and policies in ways that cannot satisfy everybody's preferences. It is true that certain policy prerogatives can be delegated to subnational levels of government through decentralization, but some policies have to be national.[13] Think for instance of defense and foreign policy, monetary policy, redistribution between regions, the legal system, etc.

The costs of heterogeneity in the population have been well documented, especially for the case in which ethnolinguistic fragmentation is used a as proxy for heterogeneity in preferences. Easterly and Levine (1997), La Porta Lopez de Silanes, Shleifer and Vishny (1999) and Alesina et al. (2003) showed that ethnolinguistic fractionalization is inversely related to economic success and various measure of quality of government, economic freedom and democracy.[14] Easterly and Levine (1997), in particular, argued that ethnic fractionalization in Africa, partly induced by absurd borders left by colonizers, is largely responsible for the economic failures of this continent. There is indeed a sense in which African borders are "wrong", not so much because there are too many or too few countries in Africa, but because borders cut across ethnic lines in often inefficient ways.[15]

We can think of trade openness as shifting the trade-off between the costs and benefits of size. As international markets become more open, the

[13] In fact, the recent move towards regional decentralization in many countries can be partly viewed as a response of the political system to increasing pressures towards separatism. See Bardhan (2002) for an excellent discussion of this point, and De Figueiredo and Weingast (2002) for a formal treatment. Also, for an excellent review of the literature on federalism, see Oates (1999).

[14] A large literature provides results along the same lines for localities within the United States. For example, see Alesina, Baqir and Easterly (1999). Related to this, Alesina and La Ferrara (2000, 2002) show that measure related to social capital are lower in more heterogeneous communities in the US. Alesina Baqir and Hoxby (2004) show how local political jurisdictions in the US are smaller in more radially heterogeneous areas.

[15] On this point see in particular Herbst (2000).

benefits of size decline relative to the costs of heterogeneity, thus the optimal size of a country declines with trade openness. Or, to put it differently, small and relatively more homogeneous countries can prosper in a world of free trade. With trade restrictions, instead, heterogeneous individuals have to share a larger polity to be economically viable. Incidentally, above and beyond the income effect, this may reduce their utility if preference homogeneity is valued in a polity. While in this paper we focus on preference heterogeneity rather than income heterogeneity, the latter plays a key role as well, a point raised by Bolton and Roland (1997). Poor regions would like to join rich regions in order to maintain redistributive flows, while richer regions may prefer to be alone. There is a limit to how much poor regions can extract due to a non-secession constraint, which is binding for the richer regions. Empirically, often more racially fragmented countries also have a more unequal distribution of income. That is, certain ethnic group are often much poorer than others and economic success and opportunities are associated with belonging to certain groups and not others. These are situations with the highest potential for political instability and violence.

2.2 A Model of Size, Trade and Growth

In this section we will present a simple model linking country size, international trade and economic growth. The model builds upon Alesina and Spolaore (1997, 2003), Alesina, Spolaore and Wacziarg (2000) and Spolaore and Wacziarg (2002).

2.2.1 Production and Trade

Consider a world in which individuals are located on a segment $[0, 1]$. The world population is normalized to 1. Each individual living at location $i \in [0, 1]$ has the following utility function:

$$\int_0^\infty \frac{C_{it}^{1-\sigma} - 1}{1 - \sigma} \, e^{-\rho t} dt \tag{1}$$

where $C_i(t)$ denotes consumption at time t, with $\sigma > 0$ and $\rho > 0$. Let $K_i(t)$ and $L_i(t)$ denote aggregate capital and labor at location i at time t. Both inputs are supplied inelastically and are not mobile. At each location i a specific intermediate input $X_i(t)$ is produced using the location-specific capital according to the linear production function:

$$X_i(t) = K_i(t) \tag{2}$$

Each location i produces $Y_i(t)$ units of the same final good $Y(t)$, according to the production function:

$$Y_i(t) = A\left(\int_0^1 X_{ij}^{\alpha}(t)dj\right)L_i^{1-\alpha}(t) \tag{3}$$

with $0 < \alpha < 1$. $X_{ij}(t)$ denotes the amount of intermediate input j used in location i at time t, and A captures total factor productivity. Intermediate inputs can be traded across different locations in perfectly competitive markets by profit-maximizing firms. Locations belong to N different countries. Country 1 includes all locations between 0 and S_1, country 2 includes all locations between S_1 and $S_1 + S_2$, .., country N includes all locations between $\sum_{n=1}^{N-1} S_n$ and 1. Hence, we will say that country 1 has size S_1, country 2 has size $S_2, ...$, country $N-1$ has size S_{N-1}, and country N has size $S_N = 1 - \sum_{n=1}^{N-1} S_n$.

Political borders impose trading costs. In particular, we make the following two assumptions:

A1). There are *no internal barriers* to trade: Intermediate inputs can be traded across locations that belong to the same country at no cost.

A2). There are barriers to international trade: If one unit of an intermediate good produced at a location within country n' is shipped to a location i'' within a different country n'', only $(1 - \beta_{n'n''})$ units of the intermediate good will arrive, where $0 \leq \beta_{n'n''} \leq 1$.

Consider an intermediate good i produced in country n'. Let $D_{in'}(t)$ denote the units of intermediate input i used domestically (i.e., either at location i or at another location within country n'). Let $F_{in''}(t)$ denote the units of input i shipped to a location within a different country $n'' \neq n'$. By assumption, only $(1 - \beta_{n'n''})F_{in''}(t)$ units will be used for production. In equilibrium, as intermediate goods markets are assumed to be perfectly competitive, each unit of input i will be sold at a price equal to its marginal product both domestically and internationally. Therefore,

$$P_i(t) = \alpha A D_{in'}^{\alpha-1}(t) = \alpha A(1 - \beta_{n'n''})^{\alpha} F_{in''}^{\alpha-1}(t) \tag{4}$$

where $P_i(t)$ is the market price of input i at time t. From equation (2) it follows that the resource constraint for each input i is:

$$S_{n'}D_{in'}(t) + \sum_{n \neq n'} S_n F_{in}(t) = K_{in'}(t) \tag{5}$$

where $S_{n'}$ is the size of country n', while $K_{in'}(t)$ is the stock of capital in location i (belonging to country n') at time t.

By substituting (4) into (5) we obtain:

$$D_{in'}(t) = \frac{K_{in'}}{S_{n'} + \sum_{n \neq n'} S_n (1 - \beta_{n'n})^{\frac{\alpha}{1-\alpha}}} \tag{6}$$

and:

$$F_{in''}(t) = \frac{(1 - \beta_{n'n''})^{\frac{\alpha}{1-\alpha}} K_{in'}}{S_{n'} + \sum_{n \neq n'} S_n (1 - \beta_{n'n})^{\frac{\alpha}{1-\alpha}}} \tag{7}$$

As one would expect, barriers to trade tend to increase the domestic use of an intermediate output and to discourage international trade.

In the rest of this analysis, for simplicity, we will assume that the barriers to trade are uniform across countries, that is:

A3). $\beta_{i'i''} = \beta$ for all i' and i'' belonging to different countries.[16]

We define:

$$\omega \equiv (1 - \beta)^{\frac{\alpha}{1-\alpha}} \tag{8}$$

This means that the lower the barriers to international trade are, the higher is ω. Hence ω can be interpreted as a measure of "international openness". ω takes on values between 0 and 1. When barriers are prohibitive ($\beta = 1$), $\omega = 0$, which means complete autarchy. By contrast, when there are no barriers to international trade ($\beta = 0$), we have $\omega = 1$, that is, complete openness.

Thus, equations (6) and (7) simplify as follows:

$$D_{in'}(t) = \frac{K_{in'}(t)}{S_{n'} + (1 - S_{n'})\omega} \tag{9}$$

and:

$$F_{in''}(t) = \frac{\omega K_{in'}(t)}{S_{n'} + (1 - S_{n'})\omega} \tag{10}$$

[16] For an analysis in which barriers are different across countries and are an endogenous function of size, see Spolaore and Wacziarg (2002).

2.2.2 Capital accumulation and growth

In each location i consumers' net household assets are identical to the stock of capital $K_{in'}(t)$. Since each unit of capital yields one unit of intermediate input i, the net return to capital is equal to the market price of intermediate input P_{it} (for simplicity, we assume no depreciation). From intertemporal optimization we have the following standard Euler equation:

$$\frac{dC_{it}}{dt}\frac{1}{C_{it}} = \frac{1}{\sigma}[P_i(t) - \rho] = \frac{1}{\sigma}\{\alpha A[\omega + (1-\omega)S_{n'}]^{1-\alpha}K_{in'}^{\alpha-1}(t) - \rho\} \quad (11)$$

Hence, the steady-state level of capital at each location i of a country of size $S_{n'}$ will be

$$K_{in'}^{ss} = \left(\frac{\alpha A}{\rho}\right)^{\frac{\alpha}{1-\alpha}}[\omega + (1-\omega)S_{n'}] \quad (12)$$

By substituting (12) into (9) and (10), and using (3), we have the following:

Proposition 1

The steady-state level of output per capita in each location i of a country of size $S_{n'}$ is

$$Y_i^{ss} = A^{\frac{1}{1-\alpha}}\left(\frac{\alpha}{\rho}\right)^{\frac{\alpha}{1-\alpha}}[\omega + (1-\omega)S_{n'}] \quad (13)$$

Hence, it follows that:

1) Output per capita in the steady-state is increasing in openness ω. That is:

$$\frac{\partial Y_i^{ss}}{\partial \omega} > 0 \quad (14)$$

2) Output per capita is increasing in country size $S_{n'}$:

$$\frac{\partial Y_i^{ss}}{\partial S_{n'}} > 0 \quad (15)$$

3) The effect of country size $S_{n'}$ is smaller the larger is ω, and the effect of openness is smaller the larger is country size $S_{n'}$. That is:

$$\frac{\partial^2 Y_i^{ss}}{\partial S_{n'}\partial \omega} < 0 \quad (16)$$

The above results show that openness and size have positive effects on economic performance, but i) openness is less important for larger countries and

ii) size matters less in a more open world.[17] In fact, were there no barriers to trade ($\omega = 1$), output would be independent of country size.

Around the steady-state, the growth rate of output can be approximated by:

$$\frac{dY}{dt}\frac{1}{Y} = \xi e^{-\xi}(\ln Y^{ss} - \ln Y(0)) \tag{17}$$

where $\xi \equiv \frac{\rho}{2}\left[(1+\frac{4(1-\alpha)}{\alpha})^{\frac{1}{2}} - 1\right]$ and $Y(0)$ is initial income.[18] Hence, we will also have:

Proposition 2

The growth rate of income per capita around the steady-state is increasing in size, increasing in openness, and decreasing in size times openness.

These results show how the economic benefits of size are decreasing in openness and the economic benefits from openness are decreasing in size. We will test the empirical implications of this model in Section 4.

2.3 The Equilibrium Size of Countries

So far we have taken the number and size of countries as given. However, in the long-run borders do change, and our model suggests that international openness may play a role in this process. As we have seen, country size affects output and growth when barriers to trade are high, while country size is less important in a world of international integration. Hence, the reduction of trade barriers should reduce the incentives to form larger countries. In what follows we will formalize this insight using the framework of country formation developed by Alesina and Spolaore (1997, 2003).[19]

If there were no costs associated with size, world welfare would be maximized by having only one country, which seems rather unrealistic. Following our previous discussion we model the costs of size as the result of heterogeneity of preferences over public policies and public goods, the collection

[17] The result does not depend on the assumption that barriers to trade are uniform across countries. In particular, one can derive analogous results for the case of non uniform barriers. Moreover, analogous results can be obtained when "openness" is defined as trade over output rather than in terms of trade barriers. See Spolaore and Wacziarg (2002).

[18] For a derivation of this result, see Barro and Sala-i-Martin (1995, chapter 2).

[19] The economics literature on the endogenous formation of political borders, while still in its infancy, has been growing substantially in the past few years. An incomplete list of contributions, besides those cited in the text, includes Friedman (1977), Casella and Feinstein (2001), Findlay (1996), and Bolton and Roland (1997).

of which we label "government". We assume that, for each location, there exists an "ideal" type of government. If individuals in location i belong to a country whose government is different from their ideal type (say $j \neq i$), their utility will be reduced by $h\Delta_{ij}$, where Δ_{ij} is the distance between j and i, and h is a parameter that measures "heterogeneity" costs - that is, the costs of being far from the median position in one's country. The distance from the government that give raise to these costs should be interpreted both as a distance in terms of preferences and in terms of location.[20]

On the other hand, in a country of size S_n the fixed costs of government can be spread through a larger population.[21] For example, if the fixed cost of government is G and it is shared equally by all citizens, each individual in a country of size S_n will have to pay G/S_n - which is obviously decreasing in S_n.

We consider the case in which borders are determined to maximize net income minus heterogeneity costs in steady-state.[22] That is, we assume that each individual at location i in a country n of size S_n is interested in maximizing the following steady-state welfare:

$$W_{in} = Y_{in}^{ss} - t_{in} - h\Delta_{in} \tag{18}$$

where Y_{in}^{ss} is steady-state income, given by $A^{\frac{1}{1-\alpha}} \left(\frac{\alpha}{\rho}\right)^{\frac{\alpha}{1-\alpha}} [\omega + (1-\omega)S_{n'}]$, t_{in} denotes taxes of individual i in country n, Δ_{in} is individual i's "distance from the government".

Country n's budget constraint is:

$$\int_{S_{n-1}}^{S_n} t_{in} di = G \tag{19}$$

How are borders going to be determined in equilibrium? First we consider how borders would be determined *efficiently*, that is, when the sum of everybody's welfare $\int_0^1 W_{in} di$ is maximized. First of all, one can immediately see

[20] This assumption is extreme but allows to have only one dimension. For more discussion see Alesina and Spolaore (2003).

[21] Obviously, not all the costs of government are fixed. Some depend positively on size, such as infrastructure spending or transfers. See Alesina and Wacziarg (1998) for an empirical examination of this point using cross-country data.

[22] The analysis could be extended in order to consider the more complex issue of border changes along the transitional dynamics, in which adjustment costs from changing borders would be explicitly modeled. Here we abstract from such issues and focus on borders in steady-state.

that the efficient solution implies countries of equal size. This is due to the assumption that people are distributed uniformly in the segment $[0,1]$.[23] Second, the government should be located "in the middle" of each country, since the median minimizes the sum of distances. When countries are all of equal size (call it $S = 1/N$, where N is the number of countries), and governments are located "in the middle", the average distance from the government is $S/4$. Hence, the sum of everybody's welfare becomes:

$$\int_0^1 W_{in} di = A^{\frac{1}{1-\alpha}} \left(\frac{\alpha}{\rho}\right)^{\frac{\alpha}{1-\alpha}} [\omega + (1-\omega)S] - \frac{G}{S} - h\frac{S}{4} \tag{20}$$

which is maximized by the following "efficient size":[24]

$$S^* = \sqrt{\frac{4G}{h - 4(1-\omega)A^{\frac{1}{1-\alpha}} \left(\frac{\alpha}{\rho}\right)^{\frac{\alpha}{1-\alpha}}}} \tag{21}$$

Hence,we have that the "efficient size" of countries is:

1) Increasing in the fixed cost of public goods provision (G),

2) Decreasing in heterogeneity costs (h),

3) Decreasing in the degree of international openness (ω),

4) Increasing in total factor productivity (A).

Therefore, in our model, if borders are set efficiently, increasing economic integration and globalization should be associated with a breakup of countries.

Should we expect such a breakup to take place if borders are *not* set optimally? For example, what if, more realistically, borders are set by self-interested governments ("Leviathans") who want to maximize their net rents? We can model the equilibrium of those Leviathans by assuming that a) they want to maximize their rents in steady-state, but) they are constrained in their rent maximization, since they must provide a minimum level of welfare to at least a fraction δ of their population (we can interpret this as a "no-insurrection constraint"). Hence, δ measures the degree to which Leviathans are constrained by their subjects' preferences.

[23] For a formal proof, see Alesina and Spolaore (1997; 2003).

[24] Equation (20) abstracts from the fact that the number of countries $N = 1/S$ must be an integer.

If we assume that each individual in a given country must pay the same taxes (that is, if we rule out inter-regional transfers), we can use t to denote taxes per person in a country of size S. Then, a Leviathan's total rents in a country of size N is given by:

$$tS - G \tag{22}$$

where t is chosen in order to satisfy the constraint:

$$W_{in} = Y_i^{ss} - t - h\Delta_i \geq W_0 \tag{23}$$

for a mass of individuals of size δS.

The Leviathan will locate the government in the middle of his country, as the social planner would do, in order to minimize the costs of satisfying (23). Constraint (23) will be binding for the individual at a distance $\delta S/2$ from the government. Hence, we have:

$$t = Y_i^{ss} - \frac{h\delta S}{2} - W_0 \tag{24}$$

By substituting (24) into (22) and maximizing with respect to S we have the following equilibrium size of countries in a world of Leviathans:

$$S^e = \sqrt{\frac{2G}{h\delta - 2(1-\omega)A^{\frac{1}{1-\alpha}}\left(\frac{\alpha}{\rho}\right)^{\frac{\alpha}{1-\alpha}}}} \tag{25}$$

Again, the size of countries is increasing in the economies of scale in the provision of public goods (G) and in the level of total factor productivity(A), while decreasing in heterogeneity costs (h) and openness (ω).

We can note that $S^e = S^*$ when the Leviathans must provide minimum welfare to exactly half of their population, while countries are inefficiently large ($S^e > S^*$) when Leviathans are really dictatorial, that is, they can stay in power without the need to take into account the welfare of a majority of the population. But even in that case, more openness induces smaller countries.

The comparative statics predict that technological progress, in a world of barriers to trade, should be associated with larger countries. This result is intuitively appealing, since technological progress improves the gains from trade, and barriers to international trade increase the importance of domestic trade, and hence a larger domestic market. However, if technological

progress is accompanied by a reduction in trade barriers, the result becomes ambiguous.[25] Moreover, a reduction in trade barriers (more openness) has a *bigger* impact (in absolute value) on the size of countries at *higher* levels of development - that is, the effect of globalization and economic integration on the size of countries is expected to be larger for more developed societies. Formally:

$$\frac{\partial^2 S^e}{\partial \omega \partial A} < 0 \tag{26}$$

Of course, these comparative statics results are based on the highly simplifying assumption that technological progress is exogenous. An interesting extension of the model would be to consider endogenous links between political borders, the degree of international openness, and technological progress.[26]

Alesina and Spolaore (2003) also analyze the case in which borders are chosen by democratic rule (majority voting). They show that in this case one may or may not obtain the efficient solution depending on the availability of credible transfer programs. When the latter are not available, in a fully democratic equilibrium in which no one can prevent border changes decided by majority rule or prevent unilateral secessions, there would be more countries than the efficient number. *A fortiori* the democratically decided number of countries would be larger that the one chosen by a Leviathan for any value of $\delta < 1$. An implication of this analysis is that democratization should lead to secessions. For the purpose of this paper, even in the case of majority rule choice of borders, the comparative statics regarding trade, size and growth are the same as in the efficient case and in the Leviathan case.

2.4 Summing up

In this section we have provided a model in which the benefits of country size go down as international economic integration increases. Conversely, the benefits of trade openness and economic integration are larger, the smaller the size of a country. Secondly, we have argued that economic integration and political disintegration should go hand in hand. As the world economy becomes more integrated, one of the benefits of large countries (the size of

[25] Another element of ambiguity would be introduced if one were to assume that the costs of government G are decreasing in A.

[26] For example, some authors have suggested that technological progress may be higher in a world with more Leviathans who compete with each other (such as Europe before and after the Industrial Revolution) than in a more centralized environment (such as China in the same period). For a recent formalization of these ideas, see Garner (2001).

markets) vanishes. As a result, the trade-off between size and heterogeneity shifts in favor of smaller and more homogeneous countries. This effect tends to be larger in more developed economies. By contrast, technological progress in a world of *high* barriers to trade should be associated with the formation of *larger* countries.

One can also think of the reverse source of causality: small countries have a particularly strong interest in maintaining free trade, since so much of their economy depends upon international markets. In fact, if openness were endogenized., one could extend our model to capture two possible worlds as equilibrium border configurations: a world of large and relatively closed economies, and one of many more smaller and more open economies. Spolaore (1995, 2001) provides explicit models with endogenous openness and multiple equilibria in the number of countries. Spolaore and Wacziarg (2002) also treat openness as an explicitly endogenous variable, and show empirically that larger countries tend to be more closed to trade. Empirically, both directions of causality between country size and trade openness, which are not mutually exclusive, likely coexist. Smaller countries do adopt more open trade policies (and are consequently more open when openness is measures using trade volumes), so that a world of small countries will tend to be more open to trade.[27] Conversely, changes in the average degree of openness in the world (brought forth for example by a reduction in trading costs) should be expected to lead to more secessions and smaller countries, as we will argue extensively below.

3 Size, Openness and Growth: Empirical Evidence

In this section, we review the empirical evidence on trade openness and growth, as well as the empirical evidence on country size and growth. We then argue that the two are fundamentally linked, because both openness and country size determine the extent of the market. Thus, their impact on growth cannot be evaluated separately. Then we estimate a specification for the determination of growth as a function of market size (itself a function of both country size and trade openness), derived directly from the model presented in Section 2. Our estimates, which are consistent with a growing body of evidence on the role of scale for growth, also provide strong support for our specific model. In particular, we show that the costs of smallness can be avoided by being open. In other words, the impact of size on growth

[27] See Alesina and Wacziarg (1998) and Spolaore and Wacziarg (2002) for cross-country empirical evidence on this point.

is decreasing in openness, or, conversely, the impact of openness on growth falls as the size of countries increases. This evidence suggests that the extent of the market is an important channel for the realization of the growth gains from trade.

3.1 Trade and Growth: A Review of the Evidence

The literature on the empirical evidence of trade and growth is vast and a comprehensive survey is beyond the scope of this article. In this subsection, we simply summarize some of the salient results from recent studies in this literature, in order to set the stage for a discussion of the more specific issue of market size and growth.

The fact that openness to trade is associated with higher growth in post-1950 cross-country data was until recently subject to little disagreement.[28] Whether openness is measured by indicators of trade policy openness (tariffs, non tariff barriers, etc.) or by the volume of trade (the ratio of imports plus exports to GDP), numerous studies document this correlation. For example, Edwards (1998) showed that, out of nine indicators of trade policy openness, eight were positively and significantly related to TFP growth in a sample of 93 countries. Dollar (1992) argued that an indicator of openness based on price deviations was positively associated with growth. Ben-David (1993) demonstrated that a sample of countries with open trade regimes displays absolute convergence in per capita income, while a sample of closed countries did not. Finally, in one of the most cited studies in this literature, Sachs and Warner (1995) classified countries using a simple dichotomous indicator of openness, and argued that "closed" countries experienced annual growth rates a full 2 percentage points below "open" countries in the period 1970-1989. They also confirmed Ben David's result: open countries tend to converge, not closed ones.

These studies focused mostly on the correlation between openness and growth, conditional on other growth determinants. In other words, little attention was typically paid to issues of reverse causation. In contrast, a more recent study by Frankel and Romer (1999) focused on trade as a causal determinant of income levels. Using geographic variables as an instrument for openness, they estimated that a 1 percentage point increase in the trade to GDP ratio causes almost a 2 percent increase in the level of per capita in-

[28] The pre-1990 literature was usefully surveyed in Edwards (1993). We will focus instead on salient papers in this literature since 1990.

come.[29] Wacziarg (2001) also addressed issues of endogeneity by estimating a simultaneous equations system where openness affects a series of channel variables which in turn affect growth. Results from this study suggest that a one standard deviation increase in the portion of the trade to GDP ratio attributable to formal trade policy barriers (tariffs, non tariff barriers, etc.) is associated with a 1 percentage point increase in annual growth across countries.

These six studies were recently scrutinized by Rodríguez and Rodrik (2000), who argued that their basic results were sensitive to small changes in specification, or that the measurement of trade policy openness captured other bad policies rather than trade impediments.[30] While it is true that cross-country empirical analysis is fraught with data pitfalls, specification problems and issues of endogeneity, these authors do recognize that it is difficult to find a specification where indicators of openness actually have a negative impact on growth.[31] In other words, they essentially conclude that the range of possible effects is bounded below by zero. One could argue that by the standards of the cross-country growth literature, this is already a huge achievement: it constitutes an important restriction on the range of possible estimates. Moreover, Rodríguez and Rodrik (2000) argue that one of the problems associated with estimating the impact of trade on growth is that protectionism is highly correlated with other growth-reducing policies, such as policies that perpetuate macroeconomic imbalances. This suggests that trade restrictions are one among a "basket" of growth-reducing policies. Since Rodriguez and Rodrik (2000), the literature on trade and growth has proceeded apace. Using a new measure of the volume of trade, Alcalá and Ciccone (2004) revisit the issue of trade and growth, and argue that "in contrast to the marginally significant and non-robust effects of trade on productivity found previously, our estimates are highly significant and robust even when we include institutional quality and geographic factors in the empirical analysis". The difference stems for these authors' use of a

[29] A crucial assumption is that the instrument (constructed as the sum of predicted bilateral trade shares, where only gravity/geographical variables are used as predictors of bilateral trade) be excludable from the growth regression, i.e. that it affects growth only through its impact on trade volumes.

[30] For another critical view of this literature, in particular of the Sachs and Warner (1995) study, see Harrison and Hanson (1999). Pritchett (1996) showed that various measures of policy openness were not highly correlated among themselves, suggesting that relying on any single measure was unlikely to capture the essence of trade policy.

[31] They state that "we know of no credible evidence–at least for the post-1945 period– that suggests that trade restrictions are systematically associated with higher growth rates.", p.317.

measure of "real openness" defined as a US dollar value of import plus export relative to GDP in PPP US dollars, as further detailed below. The same authors argue that their results are robust to controlling for institutional quality, a point disputed by Rodrik, Subramanian and Trebbi (2003). In a within-country context, Wacziarg and Welch (2003) show that episodes of trade liberalization are followed by an average increase in growth on the order of 1 to 1.5 percentage points per annum.

An important drawback of the literature on trade and growth is that it does not generally focus on the channels through which trade openness affects economic performance.[32] This makes it difficult to assess whether the dynamic effects of trade openness are mediated by the extent of the market. There are many reasons that could explain a positive estimated coefficient in a regression of trade openness (however measured) on growth or income levels. Such effects could stem from better checks on domestic policies, an improved functioning of institutions, technological transmissions that are facilitated by openness to trade, increased foreign direct investment, scale effects of the type discussed in Section 2, traditional comparative advantage-induced static gains from trade, or all of the above. Few studies attempt to discriminate between these various hypotheses. Hence, while there is a general sense that trade openness increases growth and income levels, and while this creates a presumption that market size may be important, the accumulated evidence on trade and growth does not directly answer the question of whether it is market size that is good for growth, as opposed to some other aspect of openness.

3.2 Country Size and Growth: A Review of the Evidence

We now turn to the empirical evidence on the effects of country size on economic performance. There is a vast microeconometric literature on estimating the returns to scale in economic activities and how they relate to firm or industry productivity. This literature is beyond the scope of this paper, but a general sense is that, at least in some manufacturing sectors or industries, scale effects are present. It may therefore come as a surprise that the conventional wisdom seems to be that scale effects are not easily detected at the aggregate (country) level. The macroeconomic literature on country size and growth is much smaller than the microeconometric literature, but a common claim is that the size of countries does not matter for

[32] An exception is Wacziarg (2001). Alcala and Ciccone (2004) also examine whether the effect of openness works through labor productivity or capital accumulation (in its various forms).

economic growth, either in a time-series context for individual economies, or in a cross-country context.

In a time-series context, Jones (1995a, 1995b) made a simple point. Several endogenous growth models predict that the rate of long-run growth of an economy is directly proportional to the number of researchers, itself a function of population size.[33] Hence, as the population of the United States increased (and in particular the number of scientists and researchers), so should have growth. Yet while the number of researchers exploded, rates of growth in industrial countries have been roughly constant since the 1870s. This simple empirical fact created difficulties for first-generation endogenous growth models. In particular, it was taken as indicative of the absence of scale effects in long-run growth. However, while it contributed to the conventional wisdom that scale is unrelated to aggregate growth, this finding in no way precludes the existence of scale effects when it comes to income levels, which is the focus both of the theory presented in Section 2 and of our empirical estimates presented below.[34] Hence, Jones's objection applies neither to our theory nor to our evidence. Several recent theoretical papers have sought to extend and preserve the endogenous growth paradigm while eliminating scale effects on growth. See for instance Young (1998), Howitt (1999) and Ha and Howitt (2004).

In a cross-country context, some of the most systematic empirical tests of the scale implications of endogenous growth models appeared in Backus, Kehoe and Kehoe (1992). They showed empirically, in a specification where scale was defined as the size of total GDP, that scale and aggregate growth were largely unrelated. In their baseline regression of growth on the log of total GDP, the slope coefficient was positive but statistically insignificant.[35]

[33] As suggested by Jones (1999), such models include Romer (1990), Grossman and Helpman (1991) and Aghion and Howitt (1992).

[34] Scale effects in our theory come purely from the border effect - namely the fact that it is more costly (in the iceberg cost sense) to conduct trade across borders than within. This allows us to combine scale effects with a neoclassical model of growth. Our theory has standard neoclassical implications as far as transitional growth is concerned. Thus, scale may affect growth in the transition to the steady-state, since it is a determinant of steady-state income *levels*. But scale has no impact on long-run growth, which is exogenous in our model.

[35] According to the authors, this univariate regression implies that "a hundredfold increase in total GDP is associated with an increase in per capita growth of 0.85". One could argue that this is a sizeable effect, but the t-statistic on the slope coefficient is only 1.64 and the regression contains no other control variables. In a multivariate setting, the authors show that when "standard" growth regressors (but *not* trade openness) are controlled for, the coefficient estimate on total GDP remains essentially identical, but the

Moreover, the number of scientists per countries was not found to be a significant predictor of growth, and the scale of inputs into the human accumulation process (meant to capture the extent of human capital spillovers) similarly did not help predict aggregate growth. The authors also showed that scale effects were present in the data when confining attention to the manufacturing sector (i.e. regressing manufacturing growth on total manufacturing output), and suggest that this is consistent with microeconometric studies, which typically focus on manufacturing. But the set of regressions relating to the aggregate economy is often cited as evidence that there are no effects of scale on growth at the country level.

A major problem with this approach is that variables defined at the national level may be poor proxies for the total scale of the economy, the extent of R&D activities or the importance of human capital externalities. Scale effects do not stop at the borders of countries. Since small countries adopt more open trade policies, and likely also import more technologies, a coefficient on size in a regression of growth on size that omits openness is going to be biased towards zero.[36] The authors do recognize (and show empirically) that imports of specialized inputs to production can lead to faster growth. They also mention that "by importing specialized inputs, a small country can grow as fast as a larger one". But they do not empirically examine variations in the degree of openness of an economy and how it might impact the effect of size on growth.[37] In other words, they examine separately whether country size on the one hand, and imports of specialized inputs on the other, affect growth. We propose instead to examine openness and country size jointly as determinants of market size and thus growth.

t-statistic falls considerably.

[36] See Alesina and Wacziarg (1998) and Spolaore and Wacziarg (2002) for empirical evidence that small countries tend to be more open to trade, when trade openness is measures by the trade to GDP ratio. Perhaps more surprisingly, such a relationship also holds when openness is measured by average weighted tariffs, i.e. by a direct measure of trade policy restrictiveness.

[37] Another shortcoming of the literature linking economic growth to country size is its failure to examine whether size might have different effects on growth at different levels of development. Growth may have different sources at different stages of development, and country size may affect these sources differently. For instance, scale effects may be more present in the increasing returns, endogenous growth phase that characterizes advanced industrialized countries, and have a smaller effect in the capital deepening phase that perhaps characterizes less advanced economies.

3.3 Summing up

The literature on trade and growth indicates that trade openness has favorable effects on growth and income levels, but for the most part does not inform us as to whether these effects are attributable to the extent of the market, or to other channels. The literature on scale and growth typically considers measures of scale that have to do with domestic market size (i.e. the size of a country or a national economy), and generally fails to consider that openness can substitute for a large domestic market. In what follows, we bring these literatures together to focus on the impact of market size on growth.

3.4 Trade, Size and Growth in a Cross-Section of Countries

In this subsection, we bring Propositions 1 and 2 of Section 2 to the data. If small countries tend to be more open to trade, and if trade openness is positively related to growth, then a regression of growth on country size that excludes openness will understate the effect of scale. Moreover, our theory suggests that the effects of size become less important as an economy becomes more open, i.e. the coefficient on an interaction term between openness and country size is predicted to be negative. Ades and Glaeser (1999), Alesina, Spolaore and Wacziarg (2000) and Spolaore and Wacziarg (2002) have examined how country size and openness interact in growth regressions, and have confirmed the pattern of coefficients on openness, country size and their interaction predicted by our theory. In this section, we update and expand upon these results. We focus on growth specifications of the form:

$$\begin{aligned} \log \frac{y_{it}}{y_{it-\tau}} &= \beta_0 + \beta_1 \log y_{it-\tau} + \beta_2 \log S_{it} + \beta_3 O_{it} + \beta_4 O_{it} \times \log S_{it} \\ &\quad + \beta_5' Z_{it} + \varepsilon_{it} \end{aligned} \tag{26}$$

where y_{it} denotes per capita income in country i at time t, S_{it} is a measure of country size, O_{it} is a measure of openness, and Z_{it} is a vector of control variables. In this specification, the parameter estimates on openness, country size and their interaction will be our main focus. In the context of the theory presented in Section 2, these variables as well as the Z_{it} variables are to be interpreted as determinants of the steady-state *level* of per capita income.[38]

[38] Alesina, Spolaore and Wacziarg (2000) present direct evidence on the effects of market size based on levels regressions where initial income does not appear on the right hand

3.4.1 Descriptive Statistics

Tables 1 through 3 display summary statistics for our main variables of interest, averaged over the period 1960-2000. The data on openness, investment rates, growth and income levels, government consumption, and population come from release 6.1 of the Penn World Tables (Heston, Summers and Aten, 2002), which updates their panel of PPP-comparable data to the year 2000. The rest of the data we use in this paper comes from Barro and Lee (1994, subsequently updated to 2000) or from the CIA (2002). Country size is measured by the log of total GDP or by the log of total population, in order to capture both economic size and demographic size. Throughout, we define trade openness in two ways: as the ratio of imports plus exports in current prices to GDP in current prices, and as the ratio of imports plus exports in exchange rate US$ to GDP in PPP US$. We label the first variable "nominal openness" and the second one "real openness".

Recently, Alcalá and Ciccone (2003, 2004) have criticized the widespread use of the first measure, have advocated the use of the second, finding that the latter leads to more robust effects of openness on growth. The key difference between the two measure stems from the treatment of non tradable goods. Suppose that trade openness raises productivity, but does so more in the tradable than in the nontradable sector (a plausible assumption). This will lead to a rise in the relative price of nontradables, and a fall in conventionally measured openness under the assumptions that the demand for nontradables is relatively inelastic, as it may raise the denominator of the conventional measure of openness more than the numerator. So one may observe trade-induced productivity increases going hand in hand with a decline in conventional measures of openness. "Real openness" will address the problem, since the denominator now corrects for international differences in the price of nontradable goods. We show results based on both measures, in order to simultaneously address Alcalá and Ciccone's points and to allow comparability with past results.

Table 2 reveals that both measures of openness are closely related, with a correlation of 0.87. While high, this correlation justifies examining differences in results obtained using each measure. The correlation between our two measures of country size is also high, equal to 0.85. The correlation between openness and country size is negative, whatever the measures

side. These regressions were consistent with the predictions of the theory presented in Section 2. We have repeated these levels regressions using the new cross-country data that extends to 1999, with little changes in the results.

of openness and size, and in three out of four cases is of a magnitude between 0.33 and 0.54, confirming past results that small countries are more open, and suggesting that an omission of openness in a regression of growth on country size would understate the effect of size. Finally, while the simple correlation between growth and size is 0.33 when size is measured by the log of total GDP, and the correlation between openness and growth is equal to 0.21 or 0.33 (when openness is measured in current or "real terms" respectively).

Preliminary evidence on Propositions 2 and 3 can be gleaned from conditional correlations displayed in Table 3. This table presents correlations of openness and growth conditional on country size being greater or lower than the sample median, and correlations of country size and growth conditional on openness being greater or lower than the sample median. For the sake of illustration, let us focus on the log of population as a measure of size and on current openness as a measure of openness (the results are qualitatively unchanged when using the other measures). The correlation between openness and growth is 0.51 for small countries (those smaller than 6.7 million inhabitants), and only 0.10 for large countries. Similarly, the correlation between country size and growth is 0.11 for open countries, and 0.43 for closed ones. This provides suggestive evidence that openness and country size are substitutes, and that the correlation between size and growth falls with the level of openness. To fully evaluate this claim, we now turn to panel data growth regressions.

3.4.2 Growth, Openness and Size: Panel Regressions

Tables 4 through 6 present Seemingly Unrelated Regression (SUR) estimates of regressions of growth on openness, country size and their interaction, as well as additional controls. The SUR estimator amounts to a flexible form of the random-effects panel estimator, which allows for different covariances of the error term across time periods.[39] Its use in cross-country work is now widespread (see for example Barro and Sala-i-Martin (1995)). The panel consists of four periods of 10 year-averages (1960-69, 1970-79, 1980-89 and 1990-99), and up to 113 countries. The estimation procedure is to formulate one equation per decade, constrain the coefficients to equality across periods,

[39] In contrast, the random-effects estimator imposes that the covariance between the error terms at time t and time t+1 be equal to the covariance between the error terms at time t+1 and time t+2.

and run SUR on the resulting system of equations.[40]

Table 4 present estimation results when the measure of country size is the log of population and the measure of openness involves variables in current prices. In all specifications, the parameter estimates on our three variables of interest (openness, country size and their interaction) are of the predicted sign and all are significant at the 5% level (and often at the 1% level). This holds whether we enter these variables alone (column 1), whether we control for initial income (column 2), whether we control for a long list of common growth regressors (column 3) and whether we include time specific effects in addition to all the controls (column 4). Moreover, Table 5 shows that the results change little when size is measured by the log of total GDP, although the level of significance is reduced somewhat in the specifications that include many control variables. Finally, Table 6 shows that using "real openness" does not modify the overall pattern of coefficients. In fact our results are generally stronger (in the sense of the estimated coefficients being larger in magnitude) when using this measure of openness. Similar estimates in Alcalá and Ciccone (2003, written after first draft of this paper) lend further support to our results. They show how controlling for a host of additional variables including institutional quality does not change the nature of these results and that the use of "real openness" leads to coefficients that are larger and more robust than when using "nominal openness".

3.5 Endogeneity of Openness: 3SLS estimates

Openness, especially when defined as the volume of trade divided by GDP (however deflated), may be an endogenous variable in growth regressions. As described above, in an important paper Frankel and Romer (1999) have developed a innovative instrument to deal with potential endogeneity bias in growth and income level regressions. We use our own set of geographic variables as well as Frankel and Romer's instrument to address potential endogeneity. Our panel data IV estimator relies on a three stage least squares (3SLS) procedure. This estimator achieves consistency through instrumentation, and efficiency through the estimation of cross-period error covariance terms. Table 7 presents parameter estimates of our basic specification when the list of instruments includes geographic variables, namely dummy variables for small countries, islands, small islands, landlocked countries and the

[40] We use the term constrained SUR to refer to the fact that slope coefficients are constrained to equality across periods.

interaction term between each of these measures and country size.[41] Again, the results are consistent with previous observations, namely the pattern of coefficients suggested by theory is maintained. In the specification with all the controls, the statistical significance of the coefficients of interest is reduced slightly when real openness is used instead of current openness (Table 9), though all remain significant at the 10% level. The signs of the main coefficients of interest are maintained and the magnitude of the openness coefficient is raised in all specifications, confirming the results of Alcalá and Ciccone (2003, 2004).[42]

Finally, Table 11 show the same results using the geography-based instrument from Frankel and Romer (1999), as well as the interaction term between this variable and country size. In all specifications, the signs and basic magnitudes of the coefficients of interest are unchanged (although when openness is entered in "real" terms, the estimates cease to be statistically significant at the 5% level). Spolaore and Wacziarg (2002) present more evidence on this type of regression, by treating estimating a simultaneous equations system for the endogenous determination of openness and growth jointly. Their results are similar in spirit to those presented here.

Alcalá and Ciccone (2003) present further results along the same lines, and also explicitly consider institutional quality variables in addition to performing further sensitivity tests. Their empirical results are very consistent with ours, suggesting that predictions on the relationship between trade, country size and growth implied by our model are confirmed when the "real" measure of openness is used instead of nominal openness.

3.5.1 Magnitudes and Summary

While the pattern of signs and the statistical significance of the estimates presented above is consistent with our theory, the effects could still be small in magnitude. However, they are not. To illustrate the extent of the substitutability between country size and openness, let us choose a baseline

[41] This is the same list of instruments as was used in Alesina, Spolaore and Wacziarg (2000). Unsing Hausman tests, this paper showed that this set of instruments was statistically excludable from the growth regression, and first stage F-tests suggested that they were closely related to openness and the interaction term.

[42] Tables 8 and 10 present F-tests for the first stage of the 3SLS procedure. They test the joint significant of the instruments in regressions of the endogenous variables (openness and its interaction with country size) on all the exogenous variables in the system. These F-tests show that our instruments are closely related to the variables they are instrumenting for, limiting the potential for weak instruments, especially in the specifications with many controls.

regression. Consider column 4 of Table 4 - this involves using the log of population as a measure of size, current openness as a measure of openness, and a wide range of controls in the growth regression. Consider a country with the median size. In our sample, when the data on log population are averaged over the period 1960-2000, the median country turns out to be Mali (where the log of population is 8.802 - this corresponds to an average population of 6.6 million over the sample period). The effect of a one standard deviation change in openness (a change of 42 percentage points) on Mali's annual growth is estimated to be 0.419 percentage points. In contrast, in the smallest country in our sample (the Seychelles), the same change in openness would translate into an increase in growth of 1.40 percentage points. The effect of a marginal increase in openness on growth becomes zero when the log of population is equal to 10.8, which is the size of France (in our sample, only 13 countries are larger).

Conversely, the effect of size at the median level of openness, which is attained by South Korea (with a trade to GDP ratio of 54% on average between 1960 and 1999), the effect of multiplying the country's size by 10 would be to raise annual growth by 0.33 percentage points. In contrast, a relatively closed country such as Argentina (with a trade to GDP ratio of 15% on average between 1960 and 1998) would experience an increase in growth of 0.78 percentage points from decupling its population. The effect of size on growth attains zero when openness reaches 82.4% (in our sample, 26 countries had a higher level of average openness over the 1960-1999 period). Using the results obtained with "real" measures of openness the magnitude of our results would typically be even larger.

Whether one "believes" these actual magnitudes or not, the signs and statistical significance of our variables of interest are very robust features of the data and independently confirmed and reinforced by Alcalá and Ciccone (2003). When evaluating the effects of scale on growth, it is essential to view scale as attainable either through a large domestic market, or through trade openness. Ignoring either would lead to underestimating scale effects in income. This section and the literature from which it is inspired has sought to bring together the research on the impact of trade on growth and the research on the impact of economic scale on growth, and in doing so has empirically established a substitutability between openness and country size.

4 Country Size and Trade in History

To what extent the size of countries respond to the economic "incentives" that we discussed above? Is there a sense that in the long-run the size of countries responds to economic forces? Our answer is yes, even though, of course, the determination of borders is driven by a highly complex web of politico-economic forces. The point of this section is simply to highlight the relationship between country size and trade in a brief historical excursion. We certainly we do not aim to discuss the entire history of state formation and their size. For a more extensive discussion we refer the reader to Alesina and Spolaore (2003), and to the voluminous literature cited therein.

4.1 The City-States

The city-states of Italy and the Low Countries of the Renaissance in Europe represent a clear example of a political entity that could prosper even if very small because they were taking advantage of world markets. Free trade was the key to prosperity of these small states. A contemporary observer described Amsterdam as a place were "commerce is absolutely free, absolutely nothing is forbidden to merchants, they have no rule to follow but their own interest. So when an individual seems to do in his own commercial interest something contrary to the state the state turns a blind eye and pretends not to notice".[43] The other reason why city-states could afford to be small is that the state did not provide many public goods, so that not much was lost in terms of tax burden from being small. Thus, the combination of a small states who provided very few public goods and complete freedom of trade allowed for the city state to reach unprecedented level of wealth based on trade.

4.2 The Absolutist Period

The emergence of centralized states from the consolidation of feudal manors was driven by three main forces. One is technological innovations in military technology that increased the benefits of scale in warfare. Secondly, there was a need to enforce property rights and to create markets above and beyond the maritime commerce of the city-states. Finally bellicose rulers needed vast populations in order to extract levies to finance wars and luxurious courts. Territorial expansion and fiscal pressure went hand in hand and city-states could not survive in this changed world. Italian city-states

[43] From Braudel (1992, page 206). Also cited in Alesina and Spolaore (2002).

lost predominance. The Low Countries survived longer because of their role as Atlantic traders. While the small-city states blossomed on trade, as Wilson (1967) writes regarding France "by the second half of the sixteenth century primitive ideas about trade had already given rise to a corpus of legislation ... aimed at national self-sufficiency". Similarly, English policy turned quite protectionist in the early seventeenth century. From the small and open city-states with low taxation, the western world became organized in large countries, pursuing inward looking policies. So economic predominance switched from small open economies with cheap governments to large relatively closed economies with a heavier burden of taxation to service war.

Outside the core of Europe, absolutist regimes were based on heavy taxation raised without the parenthesis of city-states. This is the case, for instance of the Ottoman Empire, but also of India and China. The Ottoman empire for instance, was largely based on extracting rents from its population. In India the level of taxation was extraordinarily high for that period. In the sixteenth century the estimated tax revenue of the central government was about 20% of GNP.

4.3 The Birth of the Modern Nation-State

The nineteenth century marks the birth of the nation-state in modern forms, both in Europe and North America. It also marks the beginning of industrialization and the growth take-off, which likely transformed the relationship between country size and economic performance, raising the importance of scale effects. The liberal philosophers of these times viewed the "optimal size" of a nation-state as emerging from the trade-off between homogeneity of language and culture and the benefit of economic size. In fact, following the work of Adam Smith, they were well aware that with free trade a market economy can easily prosper even without a heavy central government. Nevertheless, the view was that there existed an minimum size that made an economy viable. For instance, certain regions, like Belgium, Ireland and Portugal were considered too small to prosper, but free trade was regarded as a way of allowing even relatively small countries to prosper. Giuseppe Mazzini, an architect of the Italian unification, suggested that the optimal number of states in Europe was 12. His argument was precisely based on the consideration of a trade-off between the economically viable size of country and nationalistic aspiration of various groups. A famous political economy treaty of the time argues that it was "ridiculous" that Belgium and Portugal should be independent because there economies were too small to be

economically viable.[44]

The unification of Germany can in fact be viewed along similar lines. The German nation-state started as a customs union (the Zollverein) which was viewed as necessary to create a sufficiently large market. As Merriman (1996, page 629) notes, before the customs union "German merchants and manufacturers began to object to the discouraging complexity of custom tariffs that created a series of costly hurdles... many businessmen demanded an end to these unnatural impediments faced by neither of their French or British rivals". Clearly market size was a critical determinant of the birth of Germany. The external threat of a war with France was a second one, as emphasized by Riker (1964). The establishment of a common market free of trade barriers was also one of the motivating factor behind the creation of the United States.

4.4 The Colonial Empires

In the period between 1848 and early 1870's the share of international trade in GDP quadrupled in Europe.[45] From 1870 to the First World War trade grew much more slowly despite a drastic reduction of transportation costs, as documented in Estevadeordal, Frantz and Taylor (2003). In fact the extent of the reduction of trade amongst European powers in the half century between 1870 and 1915 is a matter of dispute amongst historians. Bairoch (1989) has probably the most sanguine view on one side of the argument when he writes that the introduction of new large tariff by Germany in 1879 marks the "death" of free trade. While many historians may find this view a bit extreme, it is fairly non controversial that without the sharp reduction in trading costs international trade would have probably greatly suffered in this period, which was certainly associated with an increase in protectionism.

The last two decades of the nineteenth century witnessed the expansion of European (and North American) powers over much of the "less developed" world. One motivation of this expansionary policy was certainly the opening of new markets. As reported by Hobsbawn (1987, p. 67), in 1897 the British Prime Minister told the French ambassador to Britain that "if you [the French] were not such persistent protectionists, you would not find us so keen to annex new territories". Needless to say, the British were just as protectionist as the French and the British navy was heavily used to protect trade routes. Similar considerations apply to the expansionary acquisitions

[44] See Hobsbawn (1987).

[45] See Estevadeordal, Frantz and Taylor (2002) for a more detailed discussion.

of the United States in the late nineteenth and early twentieth centuries, namely Alaska, Hawaii, Samoa, Cuba and the Philippines. At the same time, in response to European protectionism, the United States also turned protectionist in this period.

In summary, from the point of view of the colonizers, Empires were a brilliant solution to the trade-off between size and heterogeneity. Large empire guaranteed large markets, especially necessary when protectionism was on the rise, but at the same time, by not granting citizenship to the inhabitants of the colonies, the problem of having a heterogeneous population with full political rights was reduced.

4.5 Borders in the Interwar Period

Figure 1 shows all the countries created and eliminated in five years periods from 1870 until today.[46] The dip at the beginning of the figures highlights the unification of Germany. This figure shows that in the interwar period after the Treaty of Versailles, borders remained essentially frozen, despite the fact that many nationalistic aspiration had been left unanswered by the peace treaty. In fact, a common view amongst historians is that the Treaty of Versailles vastly mishandled the border issue. Nevertheless, borders remained virtually unchanged, in a period in which free trade collapsed. No decolonization occurred. Amongst the new country creations, at least one, Egypt (independent in 1922) is merely an issue of classification: it was largely independent from Britain, but its status switched from a protectorate to a semi-independent country. Leaving aside the Vatican City, the only other countries created between 1920 and the Second World War were Ireland (1921), Mongolia (1921), Iraq (1932), and Saudi Arabia (1932).

The interwar period was characterized by a collapse of free trade, the emergence of dictatorships, and by a belligerent state of international relationships. The Great Depression completed the gloomy picture and precipitated the rise of protectionism. These are all factors that, according to our analysis, should *not* be associated with the creation of new countries, in fulfillment of nationalistic aspirations. In addition, these elements (lack of democracy, international conflicts, protectionism) would make colonial powers hold on to their empires and repress independent movements. In fact, all the colonial powers were adamant in refusing self-determination of colonies during this period. This combinations of events, protectionism and

[46] This figure exclude Sub-Saharan Africa, given the difficulty of identifying borders before the colonization period.

maintenance of large countries and empire, stands in sharp contrast with what happened in the aftermath of the Second World War.

4.6 Borders in the Post-Second World War period

In the fifty years that followed the Second World War, the number of independent countries increased dramatically. There were 74 countries in 1948, 89 in 1950, and 193 in 2001. The world now comprises a large number of relatively small countries: in 1995, 87 of the countries in the world had a population of less than 5 million, 58 had a population of less than 2.5 million, and 35 less than 500 thousands. In the same 50 years, the share of international trade in world GDP increased dramatically. The volume of imports and exports in a sample of about 60 countries has risen by about 40 percent.

We should stress that the increase in international trade in the last half-century, as documented in Figure 2, is not the simple result of an accounting illusion. In fact, if two countries were to split, their resulting trade to GDP ratios would automatically increase, as former domestic trade is now counted as international trade. But Figure 2 only features the average trade to GDP ratio for a set of countries *whose borders did not change since 1870.* Furthermore, Figure 3 uses average tariffs on foreign trade for a selection of countries with available data, a more direct reflection of trade policy, to display a similar historical pattern. Obviously, such policy measures are not subject to the accounting illusion either.

The correlation between the number of countries and trade liberalization is captured by Figure 4 and 5 which plot the detrended number of independent countries against the detrended trade to GDP ratio, including Sub-Saharan Africa from 1905 onward, and without it from 1870 to 1905.[47] In both cases the correlation is very strong. Since both variables are detrended, this positive correlation is not simply due to the fact that both variables increase over time. In Figure 2, note the sharp drop in the number of countries between 1870 and 1871, due to the unification of Germany. While 1871 is on the "regression line", 1870 is well above it, suggesting that there were "too many" countries before the German unification, relative to the average level of openness.

Not only have the recent decades witnessed an increase in the number of countries, but many regions have demanded and often obtained more

[47] All these figures are take from Alesina, Spolaore and Wacziarg (2000).

autonomy from their central governments. In fact, decentralization is very popular around the world. The case of Québec is especially interesting. The push for independence in Québec was revamped by the implementation of the North American Free Trade Agreement (NAFTA). The freer trade in North America, the easier it would be for a relatively small country, like Québec, to prosper. As we discussed above, at least for Canada, national borders still matter, so that trade among Canadian provinces is much easier than trade between Canadian provinces and US states. As shown by McCallum (1988), two distant Canadian provinces trade much more with each other than US states and Canadian provinces bordering each other, even though distance is a strong determinant of trade flows. This implies that there might be a cost for Québec in terms of trade flows if it were to become independent and such arguments were made by the proponents of the "no" in the self-determination referendum of 1996. As the perceived economic costs of secession fall with greater North American economic integration, the likelihood of Québec gaining independence can be expected to increase. In fact, the development of a true free-trade area in North America might reduce these costs and make Québec separatism more attractive.

4.7 The European Union

Fifteen European countries have created a union which has several supranational institutions, such as the Parliament, a Court system, a Commission and a Council of Ministers and have delegated to them substantial policy prerogatives. We have argued that more economic integration should have lead toward political separatism. How does the European Union "fit" into this picture?

First of all, the European Union is not a state, not even a federation since it does not have the critical determinant of what a state is: the monopoly of coercion over its citizens. Thus, the European Union does not satisfy the Weberian notion of what constitutes a "sovereign state". The newly proposed draft Constitution for Europe states clearly in its article 2 that the European Union is indeed a union of independent countries and not a Federal State. Secondly, as economic integration is progressing at the European level, regional separatism is more and more vocal in several member countries of the Union, such as the UK, Spain, Belgium, Italy and even France. So much so, that many have argued that Europe will (and, perhaps should) become a collection of regions (Brittany, the Basque Region, Scotland, Catalonia, Wales, Bavaria, etc.) loosely connected within a European confederation of independent regions. In fact, ethnic and cultural

minorities feel that they would be economically "viable" in the context of a truly European common market, thus they could "safely" separate from the home country. This argument is often mentioned in the press. For an example pertaining to Scotland, see the Financial Times, September 16, 1998: "...the existence of the European Union lowers the cost of independence for small countries by providing them with a free trade area... and by creating a common currency which will relieve the Scots of the need to create one for themselves".

One way of thinking about the EU is as a supranational union of countries that have merged certain functions needed to guarantee the functioning of a common market and take advantage of economies of scale. Whether or not the attribution of responsibilities and policy prerogatives between the EU and the national government is appropriate or not is an intricate subject which is beyond the scope of this paper.[48]

5 Conclusion

This paper has argued that size matters for economic performance and that country size is endogenous and depends on economic factors such as free trade, public goods provision and preference heterogeneity. We have reviewed and extended a recent literature that has discussed country formation and secession in the context of the theory of economic growth. The econometric and historical evidence is broadly consistent with the implications of these models

Much remains to be done. On the theoretical side, we have shown how scale effects could be derived in a simple neoclassical growth model, without appealing to increasing returns technologies, endogenous R&D or human capital spillovers, but simply by appealing to the existence of a border effect driven by trading frictions. However, whether the scale effects that we observe in the data come from the border effect, technology or spillovers remains to be investigated.

The models that we discussed are based on the assumption that heterogeneity within a country has negative effects on average utility. However, heterogeneity may also bring about some benefits. In fact, the gains from trade in our model do stem from a kind of heterogeneity - the production of different intermediate goods by different regions - and this is why a larger country, for given barriers to trade, brings net economic gains through the

[48] For a discussion of this point, see Alesina and Wacziarg (1999).

trade channel. By "heterogeneity" costs here we mean the specific costs associated with disagreements over the basic characteristics of a government (including policies about official languages, religion, etc.). A richer discussion of the pros and cons of heterogeneity is certainly called for.

On the empirical side, debates are still raging. Even the literature on the effect of trade on economic performance is now subject to debates on the nature and extent of this effect. The literature on the effect of country size is even more contentious. Yet the existence of both of these effects is important to the argument that we proposed about the role of trade openness in the endogenous determination of country size. We have shown that a simultaneous consideration of an economy's openness and of its size led to estimating strong effects of both size and openness on growth in a sample of countries since 1960.

Finally, in a broad historical sweep, we have suggested that the types of trade-offs identified by our framework have been at play at various stages in modern history. In a way, current developments provide an ideal setting for observers of country creation. Since the Second World War, increasing globalization has threatened nation-states "from above", while rising regionalism and decentralizing forces have threatened them "from below". The construction of the European Union epitomizes this tension, as a fundamental redrawing of the distribution of political prerogatives is being orchestrated. Powers are being transferred down through decentralization, and up through the European construction. It is likely that if globalization proceeds apace, so will regionalism. If the backlash against globalization succeeds, however, large centralized nation-states could initiate a comeback.

References

Ades, Alberto and Edward Glaeser (1999), "Evidence on Growth, Increasing Returns and the Extent of the Market", *Quarterly Journal of Economics*, vol. 114, no. 3, August, pp. 1025-1045.

Aghion, Philippe and Peter Howitt (1992), "A Model of Growth Through Creative Destruction", *Econometrica*, March vol. 60, pp.323-351.

Aghion, Philippe and Peter Howitt (1998), "Market Structure and the Growth Process", *Review of Economic Dynamics* 1: 276-305.

Aghion, Philippe, Nicholas Bloom, Richard Blundell, Rachel Griffith and Peter Howitt (2002), "Competition and Innovation: An Inverted U Relationship", *working paper*, Harvard University, September.

Alcalá, Francisco and Antonio Ciccone (2003), "Trade, the Extent of the Market and Economic Growth 1960-1996", *unpublished*, Universitat Pompeu Fabra.

Alcalá, Francisco and Antonio Ciccone (2004), "Trade and Productivity", *Quartely Journal of Economics*, vol. 119, no. 2, May.

Alesina, Alberto, Ignazio Angeloni and Ludger Schuknecht (2001), "What Does the European Union Do?", *NBER Working Paper* #8647, December

Alesina, Alberto, Robert Barro and Silvana Tenreyro (2002), "Optimal Currency Areas", in Mark Gertler and Kenneth Rogoff, eds., *NBER Macroeconomic Annual*, vol. 17, Cambridge (MA): MIT Press.

Alesina Alberto, Reza Baqir and William Easterly (1999), "Public Goods and Ethnic Divisions", *Quarterly Journal of Economics*, vol. 114, no. 4, November, pp. 1243-1284.

Alesina Alberto, Reza Baqir, and Caroline Hoxby (2004) "Political Jurisdictions in Heterogeneous Communities" *Journal of Political Economy,* forthcoming.

Alesina Alberto and Eliana La Ferrara (2000), "Participation in Heterogeneous Communities", *Quarterly Journal of Economics*, vol. 115, no. 3, August, pp. 847-904.

Alesina Alberto and Eliana La Ferrara (2002), "Who Trusts Others?" *Journal of Public Economics*, vol. 85, August, pp. 207-34.

Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat and Romain Wacziarg (2003), "Fractionalization", *Journal of Economic Growth,* vol. 8, no. 2, June, pp. 155-194.

Alesina, Alberto, and Enrico Spolaore (1997), "On the Number and Size of Nations", *Quarterly Journal of Economics*, November, pp.1027-1056.

Alesina, Alberto, and Enrico Spolaore (2003), *The Size of Nations*, MIT Press, forthcoming.

Alesina, Alberto, Enrico Spolaore and Romain Wacziarg (2000), "Economic Integration and Political Disintegration", *American Economic Review*, vol. 90, no. 5, December, pp. 1276-1296.

Alesina, Alberto and Romain Wacziarg (1998), "Openness, Country Size and the Government", *Journal of Public Economics*, vol. 69, no 3, September, pp.305-321.

Alesina, Alberto and Romain Wacziarg (1999), "Is Europe Going Too Far?", *Carnegie-Rochester Conference Series on Public Policy*, vol. 51, no. 1, December, pp.1-42.

Bairoch, Paul (1989), "European Trade Policy, 1815-1914", in Peter Mathias and Sidney Pollard, eds, *The Cambridge Economic History of Europe*, vol. 8, 1-160. Cambridge: Cambridge University Press, 1989.

Bardhan, Pranab (2002), "Decentralization of Governance and Development", *Journal of Economic Perspectives*, vol. 16, no. 4, pp. 185-205.

Barro, Robert (1991), "Economic Growth in a Cross-Section of Countries", *Quarterly Journal of Economics*, vol. 106, no. 2, May, pp. 407-443.

Barro, Robert J. and Jong-Wha Lee (1994), "Data Set for a Panel of 138 Countries", available at http://www.nber.org/data/.

Barro, Robert J. and Xavier Sala-i-Martin (1995), *Economic Growth*, New York: McGraw Hill.

Ben-David, Dan (1993), "Equalizing Exchange: Trade Liberalization and Income Convergence", *Quarterly Journal of Economics*, vol. 108, August, pp.653-679.

Bolton, Patrick and Gerard Roland (1997), "The Breakups of Nations: A Political Economy Analysis", *Quarterly Journal of Economics*, November, pp. 1057-89

Casella, Alessandra and Jonathan S. Feinstein (2002), "Public Goods in Trade: On the Formation of Markets and Political Jurisdictions", *International Economic Review,* vol. 43, no. 2, pp. 437-462.

Central Intelligence Agency (2002), *CIA World Factbook 2002*, Dulles, VA: Brassey's Inc.

De Figueiredo, Rui and Barry Weingast (2002), "Self-Enforcing Federalism", *working paper*, Stanford University, March.

Dollar, David (1992), "Outward-oriented developing economies really do grow more rapidly: Evidence from 95 LDCs, 1976-85", *Economic Development and Cultural Change*, vol. 40, no. 3, pp.523-544.

Easterly William and Ross Levine (1997), "Africa's Growth Tragedy: Policies and Ethnic Divisions", *Quarterly Journal of Economics,* vol. 111, no. 4, November, pp. 1203-1250.

Edwards, Sebastian (1993), "Openness, Trade Liberalization and Growth in Developing Countries", *Journal of Economic Literature*, vol. 31, September, pp. 1358-1393.

Edwards, Sebastian (1998), "Openness, Productivity and Growth: What Do We Really Know?", *Economic Journal*, Vol. 108, no. 447,March, pp.383-398.

Estevadeordal, Antoni, Brian Frantz and Alan M. Taylor (2003), "The Rise and Fall of World Trade, 1870-1939", *Quarterly Journal of Economics*, vol. 118, no. 2, May.

Findlay, Ronald (1996), "Towards a Model of Territorial Expansion and the Limits of Empires" in Michelle Garfinkel and Sergios Skaperdas, eds, *The Political Economy of Conflict and Appropriation,* Cambridge UK: Cambridge University Press.

Frankel, Jeffrey A. and David Romer (1999), "Does Trade Cause Growth?", *American Economic Review*, vol. 89, no. 3, June, pp. 379-399.

Friedman, David (1977) "A Theory of the Size and Shape of Nations", *Journal of Political Economy,* vol. 85, no. 1, February, pp. 59-77.

Garner, Phillip (2001), "The Role of Rival Nation-States in Long-Run Growth", *working paper*, Brown University.

Grossman, Gene and Elhanan Helpman (1991), *Innovation and Growth in the Global Economy*, Cambridge, MA: MIT Press.

Ha, Joonkyung and Peter Howitt (2004) "Accounting for Trends in Productivity and R&D: A Schumpeterian Critique of Semi-Endogenous Growth Theory", *working paper*, Brown University.

Hall, Robert and Jones, Charles I. (1999), "Why Do Some Countries Produce so Much More Output Per Worker than Others?", *Quarterly Journal of Economics*, vol. 114 no. 1, pp. 83-116, February.

Harrison, Ann and Gordon Hanson (1999), "Who Gains From Trade Reforms? Some Remaining Puzzles", *Journal of Development Economics*, vol. 48, pp.419-447.

Helliwell, John (1998), *How Much Do National Borders Matter?*, Brookings Institution Press, Washington, DC.

Herbst, Jeffrey (2000), *States and Power in Africa*, Princeton: Princeton University Press.

Heston, Alan, Robert Summers and Bettina Aten (2002), "Penn World Table Version 6.1", Center for International Comparisons at the University of Pennsylvania (CICUP), October.

Hobsbawn, Eric (1987), *The Age of Empire*, Vintage Books, New York, NY.

Hobsbawn, Eric (1990), *Nations and Nationalism Since 1870*, Cambridge University Press, Cambridge, UK.

Howitt, Peter (1999), "Steady Endogenous Growth with Population and R & D Inputs Growing", *Journal of Political Economy*, Volume 107, Number 4, August.

Jones, Charles (1999), "Growth: With or Without Scale Effects?", *American Economic Review Papers and Proceedings*, May, Vol. 89, pp.139-144.

Jones, Charles (1995a), "R&D-Based Models of Economic Growth", *Journal of Political Economy*, August, Vol. 103, pp. 759-784.

Jones, Charles (1995b), "Time Series Tests of Endogenous Growth Models", *Quarterly Journal of Economics*, May, Vol. 110, pp. 495-525.

La Porta Rafael, Florencio, Lopez de Silanes, Andrei Shleifer and Robert Vishny (1999), "The Quality of Government", *Journal of Law, Economics and Organization*, vol. 15, no. 1, March, pp. 222-279.

Lucas, Robert E.(1988), "On the Mechanics of Economic Development", *Journal of Monetary Economics*, vol. 22, pp.3-42.

McCallum, John (1992), "On the Economic Consequences of Quebec's Separation", in A. Riggs and T. Velk (eds.), *Federalism in Peril*, Fraser University Press.

McCallum, John (1995), "National Borders Matter: Canada-US Regional Trade Patterns", *American Economic Review*, June, 615-23.

Merriman, John W. (1996), *A History of Modern Europe: From the Renaissance to the Present*, New York: W.W. Norton & Company.

Murphy, Kevin , Andrei Shleifer and Robert Vishny (1989), "Industrialization and the Big Push", *Journal of Political Economy*, vol. 87, no. 5, pp. 1003-1026.

Oates, Wallace E. (1999), "An Essay on Fiscal Federalism", *Journal of Economic Literature*, vol. 37, September, pp.1120-1149.

Obstfeld, Maurice and Kenneth Rogoff (2000), "The Six Major Puzzles in International Finance: Is There a Common Cause?", in Ben S. Bernanke and Kenneth Rogoff, eds., *NBER Macroeconomics Annual*, vol. 15, Cambridge (MA): MIT Press, pp.339-390.

Portes, Richard and Helene Rey (2000), "The Determinants of Cross-Border Equity Flows", *NBER Working Paper* No. #7336, September.

Pritchett, Lant (1996), "Measuring Outward Orientation: Can It Be Done?", *Journal of Development Economics*, vol. 49, no.2.

Riker, William H., (1964), *Federalism: Origins, Operation, Significance*, Boston: Little, Brown.

Rivera-Batiz, Luis and Paul Romer (1991), "Economic Integration and Economic Growth", *Quarterly Journal of Economics*, vol. 106, pp.531-556.

Rodrik, Dani and Francisco Rodríguez (2000), "Trade Policy and Economic Growth: A Skeptics Guide to the Cross-National Evidence", in Ben Bernanke and Kenneth Rogoff, eds., *NBER Macroeconomics Annual*, vol. 15, Cambridge (MA): MIT Press.

Romer, Paul (1986), "Increasing Returns and Long Run Growth", *Journal of Political Economy,* vol. 94, pp.1002-37.

Romer, Paul (1990), "Endogenous Technological Change", *Journal of Political Economy*, vol. 98, no. 5, October, pp. S71-S102.

Rose, Andrew (2000), "One money, one market: the effect of common currencies on trade", *Economic Policy*, vol. 15. no. 30, April.

Sachs, Jeffrey and Andrew Warner (1995), "Economic Reform and the Process of Global Integration", *Brookings Papers on Economic Activity*, no.1, pp. 1-118.

Smith, Adam (1986), *An Inquiry into the Nature and Causes of the Wealth of Nations*, Harmondsworth, UK : Penguin Books, (first published 1776).

Spolaore, Enrico (1995), "Economic Integration, Political Borders and Productivity", prepared for the CEPR-Sapir conference on "Regional Integration and Economic Growth", Tel Aviv University, December.

Spolaore Enrico (2001), "Conflict, Trade and the Size of Countries", *working paper*, Brown University.

Spolaore, Enrico and Romain Wacziarg (2002), "Borders and Growth", *NBER Working Paper* #9223, September.

Wacziarg, Romain (2001), "Measuring the Dynamic Gains from Trade", *World Bank Economic Review*, vol. 15. no. 3, October.

Wacziarg, Romain and Karen Horn Welch (2003), "Trade Liberalization and Growth: New Evidence", *NBER working paper* #10152*,* December.

Wilson, C. H. (1967), "Trade, Society and the State" in E. E. Rich and C. H. Wilson, eds., *The Cambridge Economic History of Europe from the Decline of the Roman Empire: Volume 4,* Cambridge: Cambridge University Press, pp 487-575.

Wittman, Daniel (1991), "Nations and States: Mergers and Acquisitions; Dissolution and Divorce", *American Economic Review, Papers and Proceedings*, May, pp.126-129.

Young, Alwyn (1998), "Growth without Scale Effects", *Journal of Political Economy*, vol. 106, February, pp. 41-63.

Table 1 - Descriptive Statistics (1960-2000 averages)

	No. Obs.	Mean	Standard Deviation	Minimum	Maximum
Average Annual Growth	104	1.669	1.374	-1.259	5.515
Openness Ratio (Current)	114	64.098	41.871	14.373	322.128
Openness Ratio (Real)	114	37.363	35.376	4.350	244.631
Log per capita GDP 1960	110	7.730	0.889	5.944	9.614
Log total GDP	113	23.905	1.943	19.723	29.165
Log population	114	15.763	1.678	11.019	20.670
Fertility rate	156	4.569	1.797	1.733	7.597
Female human capital	103	1.116	1.067	0.024	4.923
Male human capital	103	1.523	1.225	0.096	5.467
Investment Rate (% GDP)	114	15.653	7.880	2.023	41.252
Government consumption (% GDP)	114	19.869	9.439	4.297	48.635

Table 2 - Pairwise Correlations for the Main Variables of Interest (1960-2000 averages)

	Average Annual Growth	Log total GDP	Log per capita GDP 1960	Log Population	Openness Ratio (current)
Average Annual Growth	1.000				
Log total GDP	0.338	1.000			
Log per capita GDP 1960	0.172	0.436	1.000		
Log population	0.125	0.853	-0.058	1.000	
Openness Ratio (Current)	0.216	-0.334	0.135	-0.537	1.000
Openness Ratio (Real)	0.331	-0.042	0.382	-0.348	0.870

Table 3 - Conditional Correlations – 1960-2000

Variable	Conditioning Statement	Correlation with Growth	Number of Obs.
Openness (current)	Log pop>median=8.807	0.104	54
Openness (current)	Log pop<=median=8.807	0.511	50
Openness (current)	Log GDP> median=16.700	0.301	52
Openness (current)	Log GDP<=median=16.700	0.462	52
Openness (real)	Log pop>median=15.715	0.131	54
Openness (real)	Log pop<=median=15.715	0.579	50
Openness (real)	Log GDP> median=23.607	0.223	52
Openness (real)	Log GDP<=median=23.607	0.474	52
Log population	Openness (current)>median=53.897	0.107	50
Log population	Openness (current)<=median=53.897	0.426	54
Log GDP	Openness (current)>median=53.897	0.324	50
Log GDP	Openness (current)<=median=53.897	0.563	54
Log population	Openness (real)>median=26.025	-0.089	51
Log population	Openness (real)<=median=26.025	0.587	53
Log GDP	Openness (real)>median=26.025	0.137	51
Log GDP	Openness (real)<=median=26.025	0.625	53

Medians computed from individual samples, while correlations are common sample correlations.
Growth: Average annual growth, 1960-2000

Table 4 - Constrained SUR Estimates (size=log of population, openness=current openness)

	(1)	(2)	(3)	(4)
Size*Openness (current)	**-0.006****	**-0.006****	**-0.007****	**-0.005***
	(0.002)	**(0.002)**	**(0.002)**	**(0.002)**
Size	**0.493****	**0.481****	**0.326***	**0.412****
	(0.123)	**(0.120)**	**(0.153)**	**(0.138)**
Openness (current)	**0.057****	**0.055****	**0.059****	**0.054****
	(0.015)	**(0.014)**	**(0.020)**	**(0.018)**
Log initial per capita income		0.185	-1.157**	-1.109**
		(0.112)	(0.248)	(0.230)
Fertility			-0.332**	-0.479**
			(0.118)	(0.110)
Male human capital			0.090	0.337
			(0.279)	(0.253)
Female human capital			-0.139	-0.260
			(0.327)	(0.299)
Govt consumption (% GDP)			-0.052**	-0.035**
			(0.013)	(0.012)
Investment rate (% GDP)			0.133**	0.090**
			(0.016)	(0.016)
Intercept	-3.274**	-4.600**	8.530**	8.840**
	(1.175)	(1.355)	(3.085)	(2.84)
Intercept 1970-1979				8.170**
				(2.87)
Intercept 1980-1989				7.030*
				(2.86)
Intercept 1990-2000				6.960*
				(2.81)
# countries (# periods)	104 (4)	104 (4)	80 (4)	80 (4)
Adjusted R-squared	0.15 0.01	0.15 0.02	0.12 0.22	0.38 0.23
	0.11 0.03	0.10 0.05	0.35 0.14	0.47 0.23

Standard errors in parentheses

† significant at 10% level; * significant at 5% level; ** significant at 1% level

Table 5 - Constrained SUR Estimates (size=log of GDP, openness=current openness)

	(1)	(2)	(3)	(4)
Size*Openness (current)	**-0.005****	**-0.005****	**-0.003†**	**-0.003†**
	(0.001)	**(0.001)**	**(0.002)**	**(0.002)**
Size	**0.532****	**0.592****	**0.325***	**0.438****
	(0.099)	**(0.113)**	**(0.139)**	**(0.125)**
Openness (current)	**0.089****	**0.093****	**0.064***	**0.063***
	(0.024)	**(0.025)**	**(0.030)**	**(0.027)**
Log initial per capita income		-0.171	-1.252**	-1.342**
		(0.143)	(0.247)	(0.230)
Fertility			-0.317**	-0.466**
			(0.119)	(0.109)
Male human capital			-0.011	0.268
			(0.282)	(0.254)
Female human capital			-0.045	-0.184
			(0.331)	(0.300)
Govt consumption (% GDP)			-0.050**	-0.034**
			(0.013)	(0.012)
Investment rate (% GDP)			0.126**	0.081**
			(0.017)	(0.016)
Intercept	-8.163**	-7.937**	6.358	6.740*
	(1.758)	(1.804)	(3.471)	(3.13)
Intercept 1970-1979				6.010
				(3.16)
Intercept 1980-1989				4.820
				(3.16)
Intercept 1990-2000				4.680
				(3.12)
# countries (# periods)	104 (4)	104 (4)	80 (4)	80 (4)
Adjusted R-squared	0.11 0.01	0.12 0.01	0.13 0.22	0.41 0.24
	0.09 0.02	0.07 0.02	0.35 0.06	0.47 0.19

Standard errors in parentheses

† significant at 10% level; * significant at 5% level; ** significant at 1% level

Table 6 - Constrained SUR Estimates – Using Real Openness

	(1)	(2)	(3)	(4)
	Size=log of population	**Size=log of population**	**Size=log of GDP**	**Size=log of GDP**
Size*Real Openness	**-0.004***	**-0.006†**	**-0.008****	**-0.007***
	(0.002)	**(0.003)**	**(0.002)**	**(0.003)**
Size	**0.250****	**0.229†**	**0.496****	**0.424****
	(0.093)	**(0.129)**	**(0.096)**	**(0.126)**
Real Openness	**0.075***	**0.094†**	**0.198****	**0.185****
	(0.031)	**(0.052)**	**(0.050)**	**(0.068)**
Log per capita income, 1960	0.092	-1.295**	-0.244	-1.489**
	(0.135)	(0.235)	(0.160)	(0.238)
Fertility	-	-0.552**	-	-0.537**
		(0.111)		(0.110)
Male human capital	-	0.247	-	0.205
		(0.259)		(0.254)
Female human capital	-	-0.162	-	-0.130
		(0.298)		(0.292)
Government consumption (% GDP)	-	-0.033**	-	-0.033**
		(0.012)		(0.012)
Investment (% GDP)	-	0.090**	-	0.076**
		(0.016)		(0.017)
Intercept	-3.318	-	-8.823**	-
	(1.733)		(2.091)	
# of countries (periods)	104 (4)	80 (4)	104 (4)	80 (4)
Adjusted R-squared	-0.18 -0.01	0.33 0.21	-0.14 0.03	0.35 0.19
	-0.07 0.02	0.47 0.22	-0.03 0.06	0.50 0.24

Standard errors in parentheses

† significant at 10% level; * significant at 5% level; ** significant at 1% level

Columns (2) and (4) estimated with period specific intercepts (time effects not reported). Other specifications available upon request.

Table 7 - Constrained 3SLS Estimates (Current Openness)

	(1)	(2)	(3)	(4)	(5)	(6)
	Size=log population	**Size=log population**	**Size=log population**	**Size=log of GDP**	**Size=log of GDP**	**Size=log of GDP**
Size*Openness (current)	**-0.008** (0.002)**	**-0.007** (0.002)**	**-0.008** (0.003)**	**-0.007** (0.002)**	**-0.010** (0.002)**	**-0.003† (0.002)**
Size	**0.507** (0.157)**	**0.634** (0.144)**	**0.375* (0.176)**	**0.677** (0.143)**	**1.070** (0.167)**	**0.314* (0.158)**
Openness (current)	**0.068** (0.020)**	**0.073** (0.018)**	**0.069** (0.024)**	**0.129** (0.038)**	**0.193** (0.039)**	**0.060† (0.036)**
Log initial per capita income	-	0.147 (0.117)	-1.157** (0.251)	-	-0.525** (0.167)	-1.257** (0.247)
Fertility	-	-	-0.330** (0.120)	-	-	-0.319** (0.121)
Male human capital	-	-	0.125 (0.281)	-	-	-0.017 (0.283)
Female human capital	-	-	-0.171 (0.329)	-	-	-0.039 (0.332)
Govt consumption (% GDP)	-	-	-0.052** (0.013)	-	-	-0.050** (0.013)
Investment rate (% GDP)	-	-	0.134** (0.016)	-	-	0.126** (0.017)
Intercept	-2.701 (1.537)	-5.945** (1.513)	8.178* (3.299)	-10.843** (2.604)	-14.269** (2.561)	6.596 (3.813)
# countries (# periods)	104 (4)	104 (4)	80 (4)	104 (4)	104 (4)	80 (4)
Adjusted R-squared	0.13 0.05 0.19 0.01	0.18 –0.02 0.11 0.03	0.13 0.21 0.34 0.15	0.12 0.07 0.13 0.01	0.25 0.02 0.16 0.24	0.28 0.35 0.14 0.18

Standard errors in parentheses

† significant at 10% level; * significant at 5% level; ** significant at 1% level

Notes: Instruments used: dummies for small country, island, small island, landlocked country, and the interaction of each of these measures with the log of country size.

Table 8 - First-Stage F-Tests for the Instruments (Current Openness)

Endogenous Variable	Openness (Current)	Openness*Size
Size = log population		
Specification 1- F stat	4.83	3.92
p value	0.00	0.00
Specification 2- F stat	5.63	6.28
p value	0.00	0.00
Specification 3-F stat	4.22	4.49
p value	0.00	0.00
Size= log GDP		
Specification 4- F stat	5.61	6.25
p value	0.00	0.00
Specification 5- F stat	10.38	11.23
p value	0.00	0.00
Specification 6- F stat	7.52	7.34
p value	0.00	0.00

Note: F-tests on the instruments from a regression of each endogenous variable on the list of instruments plus the exogenous regressors in each specification.

Table 9 – Constrained 3SLS Estimates (Real Openness)

	(1)	(2)	(3)	(4)	(5)	(6)
	Size=log popula-tion	**Size=log popula-tion**	**Size=log popula-tion**	**Size=log of GDP**	**Size=log of GDP**	**Size=log of GDP**
Size*Real Openness	**-0.006* (0.003)**	**-0.006* (0.003)**	**-0.007† (0.004)**	**-0.014** (0.003)**	**-0.014** (0.003)**	**-0.007* (0.003)**
Size	**0.280** (0.107)**	**0.317** (0.103)**	**0.248† (0.146)**	**0.630** (0.111)**	**0.768** (0.124)**	**0.440** (0.141)**
Real Openness	**0.100* (0.040)**	**0.098* (0.038)**	**0.111† (0.062)**	**0.350** (0.073)**	**0.361** (0.071)**	**0.195* (0.079)**
Log per capita income, 1960	-	0.017 (0.157)	-1.277** (0.237)	-	-0.526** (0.187)	-1.493** (0.239)
Fertility	-	-	-0.543** (0.112)	-	-	-0.536** (0.110)
Male human capital	-	-	0.269 (0.260)	-	-	0.206 (0.255)
Female human capital	-	-	-0.167 (0.299)	-	-	-0.13 (0.292)
Government consumption (% GDP)	-	-	-0.033** (0.012)	-	-	-0.033** (0.012)
Investment (% GDP)	-	-	0.092** (0.017)	-	-	0.075** (0.017)
Intercept	-2.941 (1.706)	-3.922* (1.919)	-	-13.883** (2.721)	-13.503** (2.679)	-
# Countries (# periods)	104 (4)	104 (4)	80 (4)	104 (4)	104 (4)	80 (4)
Adjusted R-squared	-0.17 -0.01 –0.09 0.01	-0.20 -0.01 -0.06 0.00	0.33 0.22 0.46 0.22	-0.10 0.02 -0.15 -0.01	-.21 -0.01 - 0.08 -0.02	0.35 0.19 0.50 0.24

Standard errors in parentheses

† significant at 10% level; * significant at 5% level; ** significant at 1% level

Notes: Instruments used: dummies for small country, island, small island, landlocked country, and the interaction of each of these measures with the log of population.

Columns (3) and (6) estimated with period specific intercepts (time effects not reported). Other specifications available upon request.

Table 10 – First-Stage F-Tests for the Instruments (Real Openness)

Endogenous Variable	Openness (Constant)	Openness*Size
Size= log GDP		
Specification 1- F stat	4.45	4.95
p value	0.00	0.00
Specification 2- F stat	9.09	9.92
p value	0.00	0.00
Specification 3- F stat	10.75	10.80
p value	0.00	0.00
Size = log population		
Specification 4- F stat	4.55	3.52
p value	0.00	0.00
Specification 5- F stat	6.25	7.18
p value	0.00	0.00
Specification 6- F stat	5.67	7.20
p value	0.00	0.00

Note: F-tests on the instruments from a regression of each endogenous variable on the list of instruments plus the exogenous regressors in each specification.

Table 11 – Constrained 3SLS Estimates (using Frankel and Romer's Instrument)

	(1)	(2)	(3)	(4)
	Size=log of population	**Size=log of population**	**Size=log of GDP**	**Size=log of GDP**
	Current Openness	**Real Openness**	**Current Openness**	**Real Openness**
Size*Openness	**-0.008****	**-0.010†**	**-0.003†**	**-0.009***
	(0.003)	**(0.006)**	**(0.002)**	**(0.004)**
Size	**0.435***	**0.273**	**0.399***	**0.452****
	(0.180)	**(0.197)**	**(0.166)**	**(0.173)**
Openness	**0.128****	**0.163†**	**0.089†**	**0.242***
	(0.041)	**(0.088)**	**(0.049)**	**(0.099)**
Log initial per capita income	-1.114**	-1.254**	-1.282**	-1.433**
	(0.251)	(0.252)	(0.245)	(0.255)
Fertility	-0.307*	-0.354**	-0.290*	-0.348**
	(0.122)	(0.120)	(0.125)	(0.118)
Male human Capital	0.105	-0.011	-0.036	-0.086
	(0.280)	(0.291)	(0.283)	(0.284)
Female human Capital	-0.164	-0.023	-0.043	0.031
	(0.321)	(0.327)	(0.325)	(0.320)
Government consumption (% GDP)	-0.053**	-0.052**	-0.051**	-0.052**
	(0.013)	(0.013)	(0.013)	(0.013)
Investment rate (% GDP)	0.131**	0.130**	0.122**	0.112**
	(0.017)	(0.017)	(0.017)	(0.019)
Intercept	3.959	7.991	2.219	2.694
	(4.408)	(4.296)	(4.948)	(4.547)
# countries (# periods)	80 (4)	78 (4)	80 (4)	80 (4)
Adjusted R-squared	0.12 0.21 0.36 0.14	0.04 0.20 0.37 0.10	0.11 0.23 0.37 0.02	0.02 0.18 0.40 0.12

Standard errors in parentheses

† significant at 1% level; * significant at 5% level; ** significant at 1% level

Notes: Instruments used: Frankel-Romer instrument for openness and its interaction with the log of GDP.

Figure 1. Countries Created and Destroyed
(5-year periods, excludes Sub-Saharan Africa)

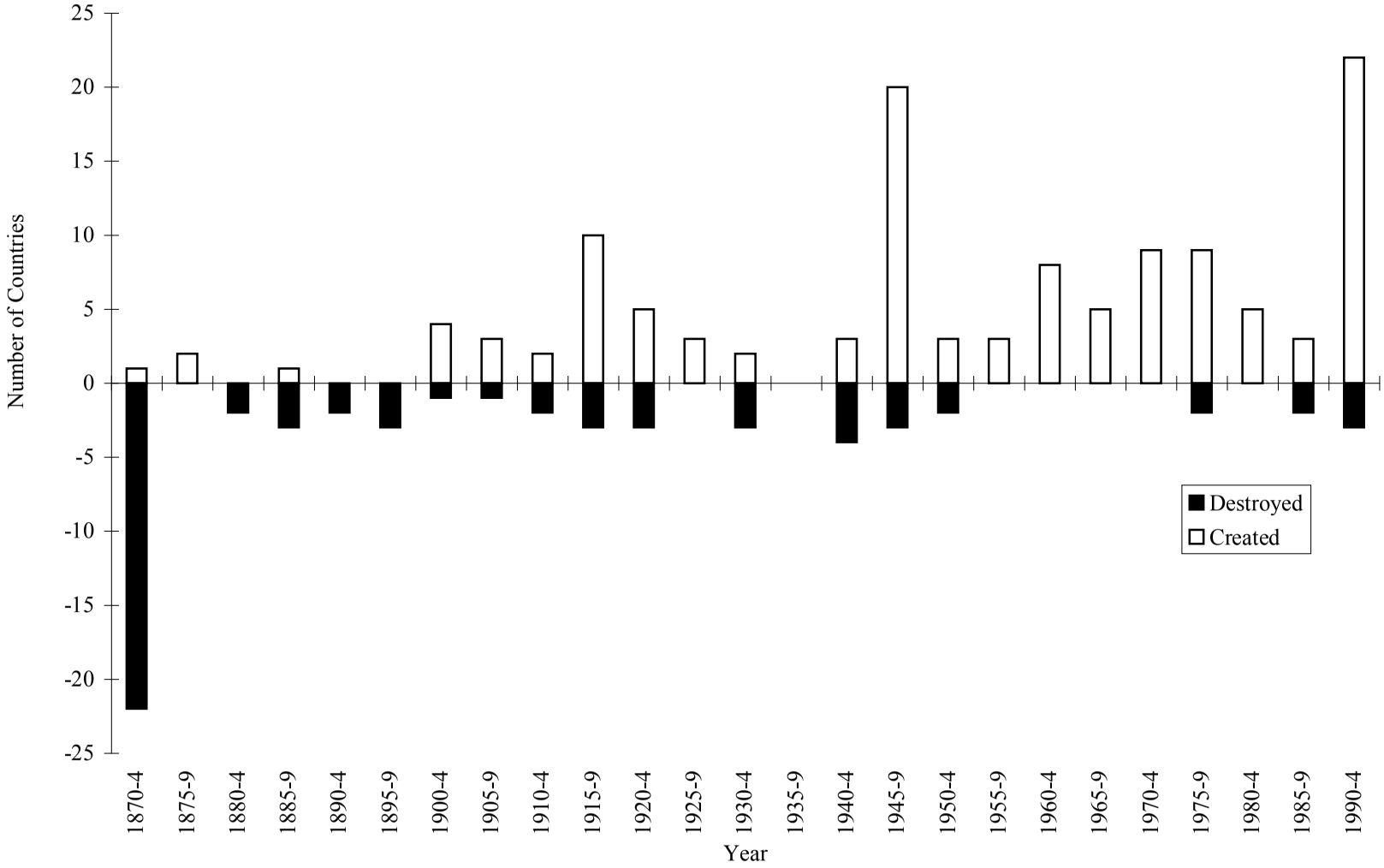


Figure 2. Trade Openness and the Number of Countries

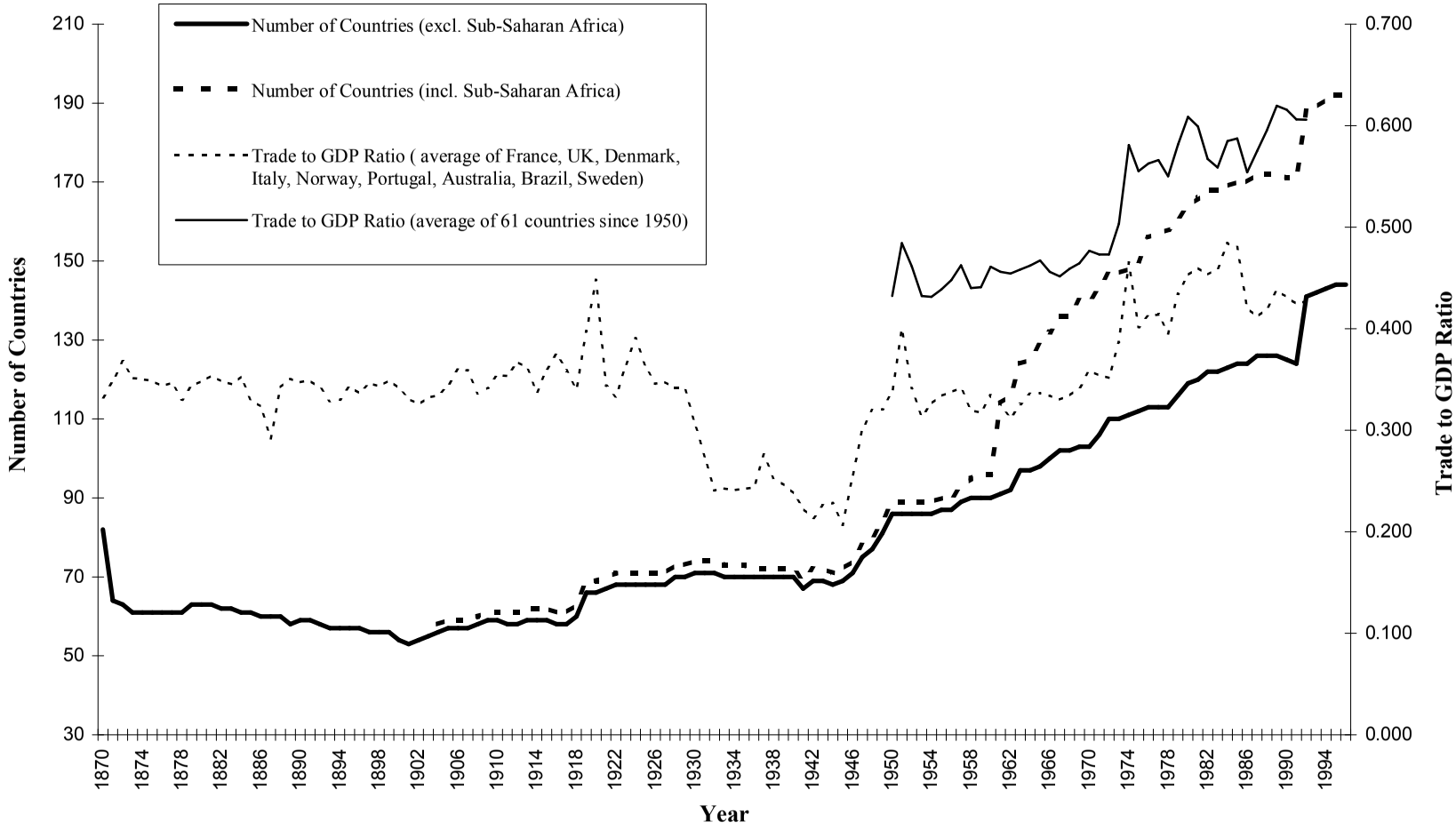


Figure 3: Average Tariff Rate and the Number of Countries
(Unweighted country average of average tariff rate for Austria, Belgium, France, Germany, Sweden, USA)

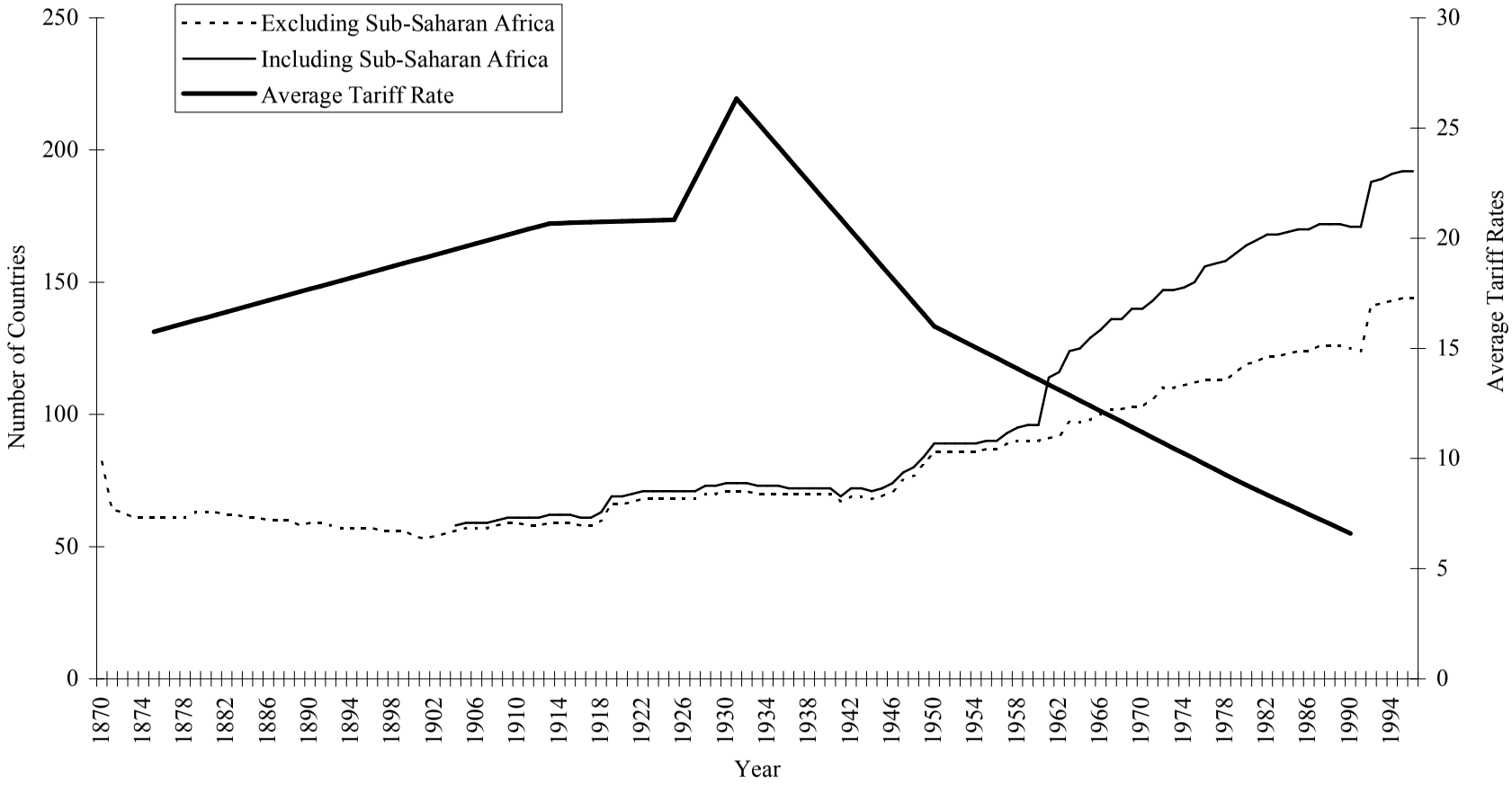


Figure 4. Scatterplot of the Detrended Number of Countries Plotted Against the Detrended Trade to GDP ratio (Without Sub-Saharan Africa - 1870–1992)

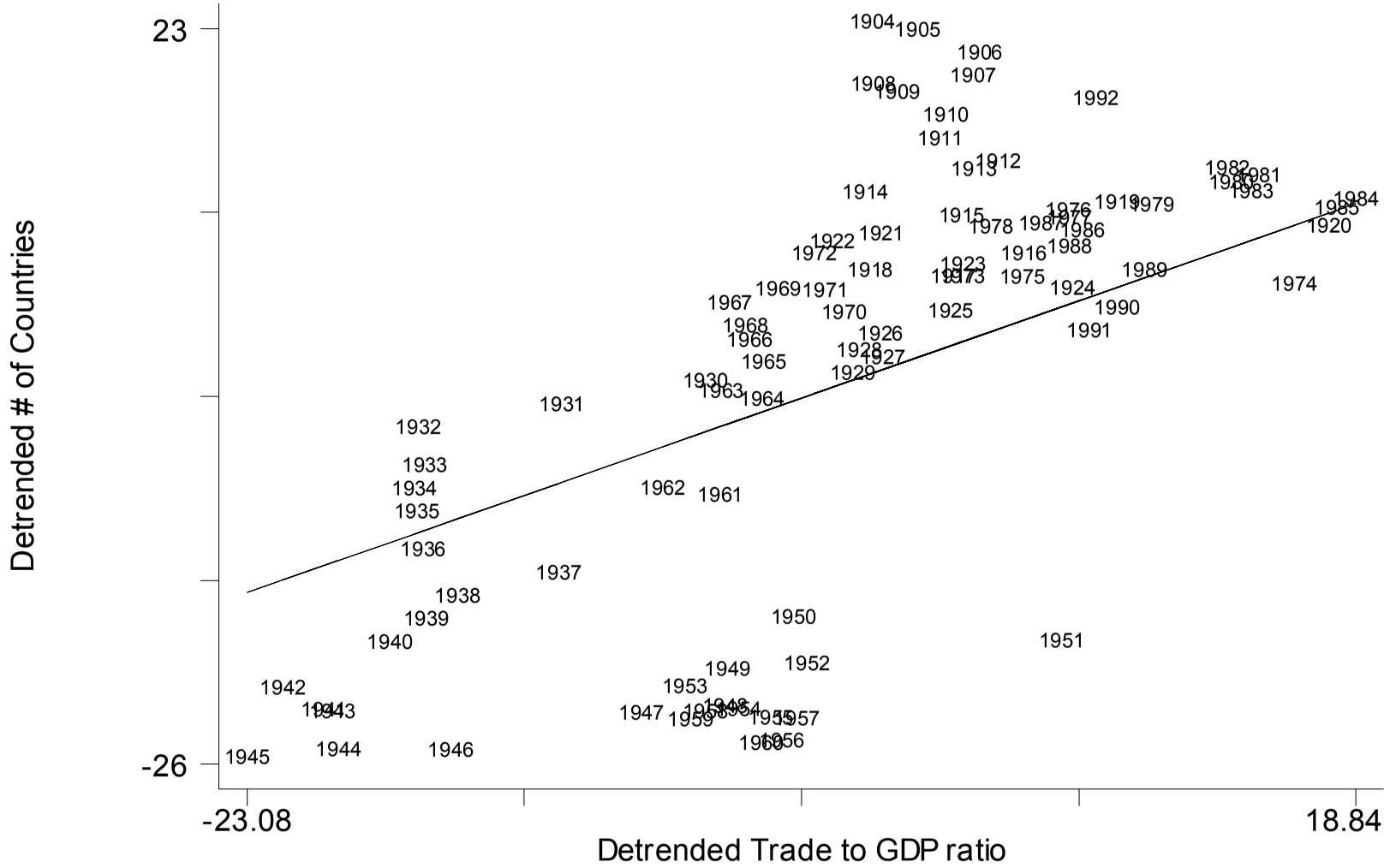
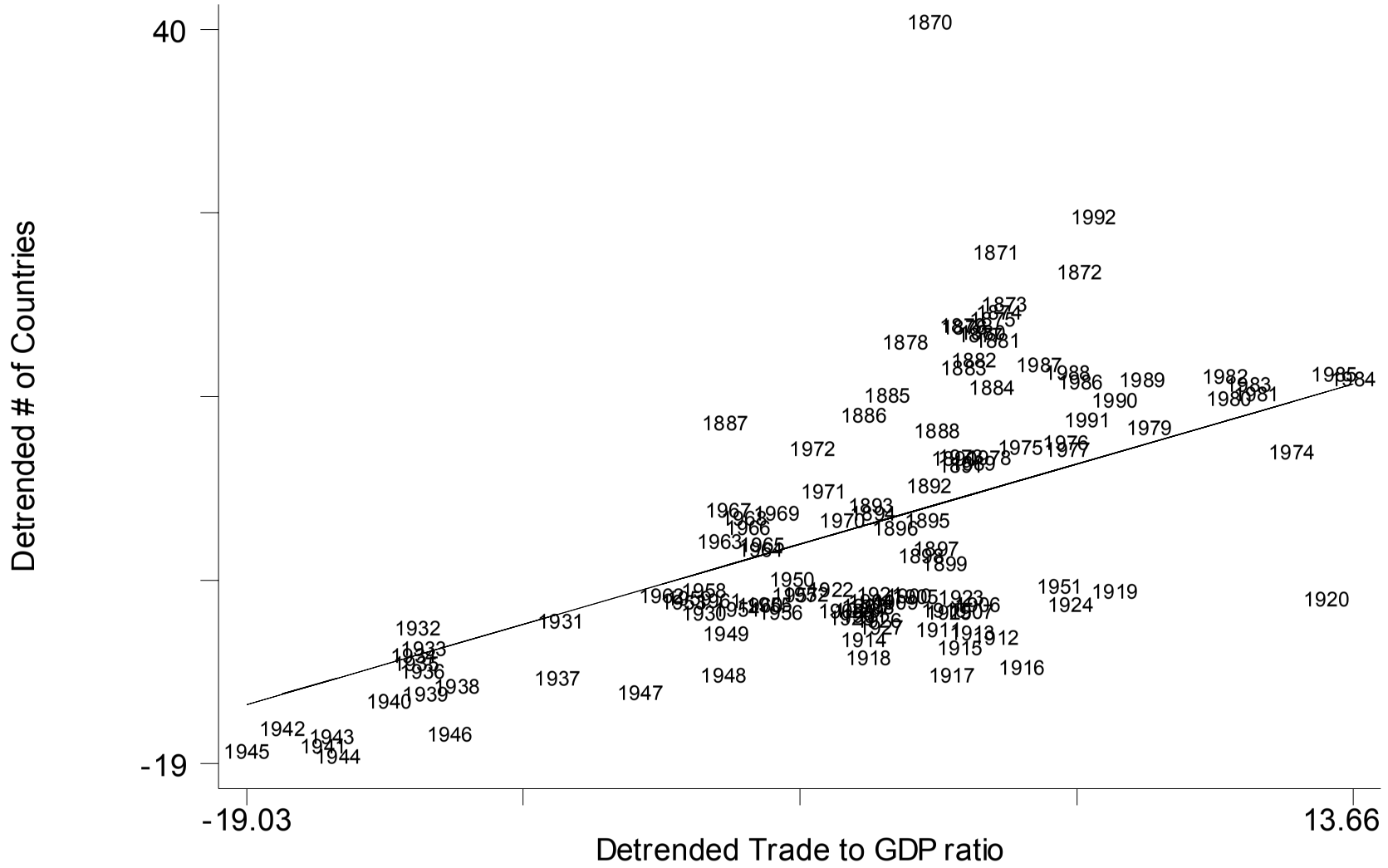


Figure 5. Scatterplot of the Detrended Number of Countries Plotted Against the Detrended Trade to GDP ratio (With Sub-Saharan Africa - 1903–1992)



Urbanization and Growth[*†]

J. Vernon Henderson

Brown University

August 11, 2004

[*] I thank Diego Puga for very helpful comments on a preliminary draft of this chapter.

[†] Draft chapter prepared for Handbook of Economic Growth, Volume 1, P. Aghion and S. Durlauf (eds.), North Holland.

The study of urbanization and growth focuses on a set of five inter-linked sets of questions. First, why do cities form and why is economic activity so geographically concentrated? In the USA, only 2% of the land area is covered by the urban built environment. This incredible geographic concentration is the central focus of economic geographers. Economists dating from Marshall (1890) have answered the question by saying urban agglomerations are based on technological externalities – the information spill-over benefits in input and output markets of having economic agents in close spatial proximity, where information decay over space is very rapid. In addition the new economic geography develops the idea that close spatial proximity involves pecuniary externalities – reduces the costs of intermediate and final good trade. Agglomeration benefits are specified typically as applying within industries or sets of inter-related industries; there is considerable debate empirically about their application across industries. That issue, as we will see later, is related to the second set of questions.

How do cities interact with each other, at any instant in time? What are the trade patterns across cities in final and intermediate outputs and how does that correspond to the roles of big and small cities? In what ways are cities specialized by either products or functions, and why? How do these patterns of specialization and diversification relate to city labor force compositions and human capital accumulations?

Given the role of cities at a point in time, the third set of questions asks how urban growth intersects with, or even defines national economic growth? The close connection between urban and national economic growth was recognized by Lucas (1988) and inspired by the development of endogenous growth models. To the extent endogenous growth is based on knowledge spillovers and sharing, given the role of close spatial proximity in spillovers, much of the interaction and sharing must occur at the level of individual cities. Given that, there must also be a close connection between economic development and urbanization. How are the two tied together? In addition the stochastic forces that shock production processes, invention, and technological progress must also play out in an urban form. How does that occur?

The fourth set of questions asks how governance, institutions, and public policy affect urbanization, which then in turn affect economic efficiency and growth. Apart from the long standing analysis of provision and financing of local public goods, there are three issues of interest specific to the urbanization process. First, public infrastructure investments in cities are enormous and the internal structure of cities affects not just the resources devoted to urban living such as commuting and congestion costs, but also affects production efficiency – the extent to which information and knowledge spillovers are fully realized and exploited. Second, institutions governing land markets, property rights, local government autonomy, and local financing including local public debt accumulation affect the city formation process, city sizes, and national economic growth. Finally, national government policies

concerning migration, trade policy, national investment in communications and transport infrastructure have profound impacts on the urban system, migration patterns, regional economic development and the like.

The final set of questions has to do with where cities locate and the economic geography of urbanization. In what regions do cities cluster and why are some regions so sparsely populated? What first nature forces of location of natural resources, including rivers and natural harbors drive the location of economic activity? How do transport costs and technological change in transport costs affect the extent to which coastal versus hinterland regions are inhabited? And what is the role of second nature forces and history on location – how does the accumulation of economic activity based on historical market forces affect the current spatial patterns of economic activity?

This handbook chapter reviews evidence on all these questions and then turns to models that focus on aspects of the middle three questions – how do cities interact with each other; what is the relationship among urbanization, urban growth, national economic growth, and economic development; and what is the role of institutions and public policy in shaping urbanization? In terms of the first question on why cities form, there is a splendid handbook paper by Duranton and Puga (2004) reviewing models of the micro-foundations of agglomeration economies and another by Rosenthal and Strange (2004) reviewing empirical evidence on the subject. In terms of the where question, there is little in the way of models that look at the location patterns of individual cities. There are the core-periphery models of economic geography that analyze the allocation of economic activity within a country between a core and periphery region. We will discuss how these models may inform the where question for cities. But they are a topic unto themselves with excellent general handbook coverage in Overman, Venables, and Redding (2001) and coverage specific to regional issues in Ottaviano and Thisse (2004), with a review of empirical evidence in Head and Mayer (2004).

The first section reviews data and empirical evidence on aspects of the five questions. The second presents a simple system of cities model, which illustrates the basic organization of the urban sector and the interaction between economic and urban growth. The model serves as a platform to discuss issues of institutions and policy. In the third section, the model in Section 2 is adapted to analyze rural-urban transformation and urbanization as part of economic development; policy issues for developing countries are analyzed.

1. Facts and Empirical Evidence

This section reviews basic facts and a body of empirical evidence on systems of cities. We start by looking at evidence based primarily on either the world as a whole or on large developed countries.

We look at the evolution of the size distribution of cities, Zipf's Law and related topics. Then we turn to what cities do – evidence on urban specialization and geographic concentration – and where they locate. Finally we turn to evidence that is more specific to the urbanization process in developing countries and issue surrounding that process.

1.1 The Size Distribution of Cities and Its Evolution

Work by Eaton and Eckstein (1997) on France and Japan and by Dobkins and Ioannides (2001) on the USA, with later work by Black and Henderson (2003) and Ioannides and Overman (2003) on the USA, establish some basic facts about urban systems and their development in France, Japan, and the USA over the last century or so. Foremost is that there is a wide relative size distribution of cities in large economies that is stable over time. Big and small cities coexist in equal proportions over long periods of time. Second, within that relative size distribution, individual cities are generally growing in population size over time; and what is considered a big versus small city in absolute size changes over time. Third, while there is entry of new cities and rapid growth and relative decline of cities nearer the bottom of the urban hierarchy, at the top city size rankings are remarkably stable over time. Finally, size distributions of cities within countries, at least at the upper tail are well approximated by a Pareto distribution, with Zipf's Law applying in many cases. Establishing these facts raises a variety of issues and different methodological and technical approaches.

1.1.1 What is a City?

The empirical work in Eaton and Eckstein (1997) and subsequent work typically looks at the decade by decade development of urban systems. In doing so, there are critical choices researchers must make when assembling data. First is to define geographically what consists of the generic term "city". The usual definition is the "metro area", where large metro areas like Chicago comprise over 100 municipalities, or local political units. The idea in defining metro areas is to cover the entire local labor market and all contiguous manufacturing, service and residential activities radiating out from the core city, until activity peters out into farm land or very low density development. A second choice concerns how to accommodate changes in geographic definitions over time. One can use whatever contemporaneous definitions the country census/statistical bureau uses; however metro area definitions only start to be applied after World War II. Another approach is to take current metro area definitions and follow the same geographic areas back in time, focusing on non-agricultural activity.

A third problem concerns how to define "consistently" over time the threshold population size at which an agglomeration becomes a metro area, especially since the economic nature, population density, and spatial development of metro areas have changed so much over the last century. Some authors use an absolute cut-off point (e.g., urban population of 50,000 or more); some use a relative cut-off point (e.g., the minimum size city included in the sample should be .15 mean city size); and others

look at a set number (e.g., 50 or 100) of the largest cities. The relative cut-off point approach is attractive because it attempts to hold constant the area of the relative size distribution which is examined over time, as illustrated below. In presenting evidence on the topics to follow, whatever choices researchers make can strongly affect specific results. Nevertheless there are a variety of findings that are consistent across studies.

1.1.2 Evolution of the Size Distribution

In the research, one focus has been to study the evolution of the size distribution of cities, applying techniques utilized by Quah (1993) in examining cross-country growth patterns. Cities in each decade are divided by relative size into, say, 5-6 discrete categories, with fixed relative size cut-off points for each cell (e.g., < .22 of mean size, .22 to.47 of mean size, ... > 2.2 mean size). A first order Markov process is assumed and a transition matrix calculated. In many cases, stationarity of the matrix over decades can't be rejected, so cell transition probabilities are based on all transitions over time. If M is the transition matrix, i the average rate of entry of new cities in each decade (in a context where in practice there is no exit), Z the (stationary) distribution across cells of entrants (typically concentrated on the lowest cell), and f the steady-state distribution, then

$$f = \left[I - (1-i)\ M \right]^{-1}\ iZ \qquad (1)$$

In the data, relative size distributions are remarkably stable over time and steady-state distributions tend to be close to the most recent distributions. In the studies on the USA, Japan, and France, there is no tendency of distributions to collapse and concentrate in one cell, or for all cities to converge to mean size; nor generally is there a tendency for distributions to become bipolar. Distributions are remarkably stable. I illustrate this based on a world cities analysis (although conceptually distributions may better apply to "countries", within which population is relatively mobile).

Table 1 gives the size distribution of world metro areas over 100,000 population in 2000. Details on the data are available on-line.[1] Note that much of the world's population in cities over 100,000 are in small-medium size metro areas. 56% are in cities under 2 million, while only 17% are in cities over 8 million. Moreover all these cities only account for 62% of the world's urban population; the rest live in cities smaller than 100,000. So overall 73% of the world's urban population lives in cities under 2 million in population. While the popular press may focus on mega-cities, only a small part of the action is there.

Figure 1 plots the relative size distribution of the approximately 1200 metro areas worldwide over 100,000 in 1960 against the relative size distribution of the approximately 1700 metro areas over 200,000

[1] http://www.econ.brown.edu/faculty/henderson/worldcities.html.

in 2000. Relative sizes are actual sizes divided by the world average size in the corresponding year. The 100, 000 versus 200,000 cut-off points for minimum size are relative ones based on a constant minimum to mean size ratio. Using an absolute cut-off point in this case has little impact on the figure. The figure plots the histogram for 20 cells on a log scale. The 1960 versus 2000 distributions for all cities worldwide (Figure 1a) and for those in developing and transition economies (Figure 1b) almost perfectly overlap. Relative size distributions are stable. Similarly performing transition analysis on world cities for 1960-1970-1980-1990-2000 and calculating the steady state distributions, starting with 5 cells and shares in each of .351, .299, .151, .100, .0991 in 1960, as we move up the urban hierarchy the steady state shares are .324, .299, .138, .122, and .117. Again, this indicates rock stablility of distributions over time.

An alternative way of expressing this is to calculate spatial Gini's (Krugman, 1991b). For a spatial Gini rank all cities from smallest to largest on the x-axis and on the y-axis calculate their Lorenz curve -- the cumulative share of total sample population. The Gini is the share of the area below the 45° line, between that line and the Lorenz curve. The greater the Gini, the "less equal" the size distribution. The world Gini in 1960 versus 2000 is .59 versus .56 for developed countries, .57 versus .56 for less developed countries, and .52 versus .45 for transition economies as noted in Table 2 columns 1-4. Table 2 also lists Gini's for 1960 versus 2000 for 14 countries. Note apart from transition economies (and Nigeria), the lack of change; and note also that transition economies are distinctly "more equal". Transition economies have forestalled the growth of mega-cities through explicit and implicit (housing availability in cities) migration restrictions, as discussed in Section 3.3.1.

A second finding in examining city size distributions is that, for larger cities, over time there is little change in relative size rankings. In Japan and France, the 39-40 largest cities in 1925 and 1876 respectively all remain in the top 50 in 1985 and 1990 respectively; and, at the top, absolute rankings are unchanged (Eaton and Eckstein, 1997). The USA displays more mobility due to substantial entry of new cities. However, while smaller cities do move up and down in rank, the biggest cities tend to remain big over time. So, for example, cities in the top decile of ranking stay in that decile indefinitely, with newer cities joining that decile as the total number of cities expands. Alternatively viewed, based on the Markov transition process, the mean first passage time for a city to move from the top to bottom cell is thousands of years (Black and Henderson, 2003). In the world cities data, as in the USA data, the probability in the transition matrix of moving out of the top cell to the next cell is very small: .038 in a decade time frame. Why do big cities stay big? A common answer is physical infrastructure (see Section 3.3.2). Large cities have huge historical capital stocks of streets, buildings, sewers, water mains and parks that are cheaply maintained and almost infinitely lived in, that give them a persistent comparative advantage over cities without that built-up stock. A second answer is modeled in Arthur (1990) and Rauch (1993) where, with localized scale externalities in production, large cities with a particular set of industries have a

comparative advantage in attracting new firms, relative to cities with a small representation of those industries. Large cities have an established scale, offering high levels of scale externalities, which smaller cities can only achieve quickly if they are able to co-ordinate mass in-migration of firms into their location, something which may be institutionally difficult to do.

1.1.3 Growth in City Numbers and Sizes

For any steady state size distribution of cities, as urbanization and growth proceed, both the absolute sizes and numbers of cities have grown historically, as a country's urban population expands through rural-urban migration and overall population growth. City sizes in the USA, Japan, and France over the past century have grown at average annual rates of 1.2 - 1.5%, depending on the country and exact time interval, rates which involve city sizes rising 3.3 - 4.5 fold every century. A small city today which is 250,000 would have been a major center in 1900.

In the world city data set, for comparable sets of countries the numbers of metro areas grew by 62% from 1960-2000 using a relative cut-off point (approximately 100,000 in 1960 versus 170,000 for this sample in 2000). Average sizes grew by about 70%. Decade by decade figures are given in Table 3. Using an absolute cut-off point of 100,000, numbers have about doubled and average sizes grown by 36% over 40 years. However we count cities, it is clear they have grown in population and numbers on an on-going basis over the decades.

The theory section will model city size growth and numbers in developed, or fully urbanized countries in Section 2 and in urbanizing economies in Section 3, as related to technological change induced by knowledge accumulation and demographic changes. There is empirical work relating city size increases to changes in knowledge levels. Glaeser, Scheinkman, and Shleifer (1995) in a cross-section city growth framework estimate that controlling for 1960 population, in the USA, cities in 1990 are 7% larger if they had a one-standard deviation higher level of median years of schooling in 1960. Black and Henderson (1999) place the issue in a panel context for 1940-1990 for the USA controlling for city fixed effects and examining the impact of percent college educated (which has enormous time variation). They find a one-standard deviation higher level of percent college educated leads to a 20% larger city.

1.1.4 Zipf's Law

In considering the size distribution of cities, especially in a cross-sectional context, there is a large literature on what is termed Zipf's Law (e.g., Rosen and Resnick 1980, Clark and Stabler 1991, Mills and Hamilton 1994, and Ioannides and Overman 2003). City sizes are postulated to follow a Pareto distribution, where if R is rank from smallest, r, to largest, 1, and n is size

$$R(n) = An^{-a} \tag{2}$$

given the Prob $(\tilde{n} > n) = An^{-a}$ and relative rank is $R(n)/r$, or the proportion of cities with size greater than n. Under Zipf's Law $a = 1$, or we have the rank size rule where, for every city, rank times size is a constant, A. Putting (2) in log-linear form, empirical work produces a's that vary across countries, samples, and times but many are "close" to one. This empirical regularity has drawn considerable attention and is often used to characterize spatial inequality, using (2) as a first approximation of the true size distribution. We list sample a coefficients for 2000 for fifteen countries in Table 2, column 5. Note however while people often say that an exponent of .74 or 1.34 is "close to" one, such coefficients produce very different city size distributions, than if the coefficient is one.[2] As a declines, or the slope of the rank size line gets flatter, urban concentration is viewed as increasing: for given size changes, rank changes more slowly, or cities are "less equal". In Table 2, the a coefficients and the Ginis are in fact strongly negatively correlated, as one would expect. But we note that typically the log version of equation (2) is better approximated by a quadratic form than linear one. However one looks at it, Zipf's Law is just an approximation that does well in some circumstances and not so well in others. If one wants to compare measures of urban concentration across countries, rather than compare estimated *a*'s in eq. (2), a more non-parametric measure such as a Gini might be more reliable.

If Zipf's Law holds even approximately, why is that? In an interesting development, Gabaix (1999a, 1999b) starts to formalize the underlying stochastic components which might lead to such a relationship, building on Simon (1955). Gabaix shows that if city growth rates obey Gabaix's Law where growth rates are random draws from the same distribution,[3] so growth rates are independent of current size, Zipf's Law emerges as the limiting size distribution. Growth is scale invariant, so the final distribution is; and we have a power law with exponent 1. Gabaix sketches an illustrative model, based on on-going natural amenity shocks facing cities of any size, which leads to Zipf's Law for the size distribution of cities. More comprehensive formulations in Duranton (2004) and Rossi-Hansberg and Wright (2004) are discussed in the theory section.

While Gibrat's Law is a neat underlying stochastic process, does it hold up empirically? Black and Henderson (2003) test whether in the relationship, $\ln n_{it} - \ln n_{it-1} = a + \delta t + \alpha \ln n_{it-1} + \varepsilon_{it}$, $\alpha = 0$ as hypothesized under the Law. The Law requires ε_{it} to be i.i.d., so simple OLS suffices. Black and Henderson find $\alpha < 0$ under a variety of circumstance and sub-samples, under appropriate statistical criteria, which rejects Gibrat's Law. Ioannides and Overman (2003) examine the issue more thoroughly in a non-parametric fashion, characterizing the mean and variance of the distribution from which growth

[2] I don't report standard errors since OLS estimates of standard errors are biased downwards. See Gabaix and Ioannides (2004).

[3] Actually the requirement is that they face the same mean and variance in the drawing.

rates are drawn. The mean and variance of growth rates do seem to vary with city size but bootstrapped confidence intervals are fairly wide generally, allowing for the possibility of (almost) equal means.

1.2 Geographic Concentration and Urban Specialization

Geographic concentration refers to the extent to which an industry k is concentrated at a particular location or, more generally concentrated at a few versus many locations nationally. A common measure of concentration of industry k at location i is $l_{ik} = X_{ik} / \sum_i X_{ik}$. X_{ik} is location i 's employment or output of industry k. Thus l_{ik} is location i 's share of, say, national employment in industry k. In contrast to geographic concentration, specialization refers to how much of a location's total employment is found in industry k, or $s_{ik} = X_{ik} / \sum_k X_{ik}$. As Overman, Redding and Venables (2001) demonstrate, if we normalize l_{ik} by location i 's share of national employment $(s_{ik} \equiv \sum_k X_{ik} / \sum_k \sum_i X_{ik})$ and s_{ik} by industry k 's share of national employment $(s_k \equiv \sum_i X_{ik} / \sum_k \sum_i X_{ik})$ we get the same measure -- a location quotient, or

$$q_{ik}^k = X_{ik} \frac{(\sum_k \sum_i X_{ik})}{\sum_k X_{ik} \sum_i X_{ik}} \tag{3}$$

The distribution of q_{ik} across industries, k, compared over time for a city would tell us about how city i 's specialization patterns are changing over time. And the distribution of q_{ik} across locations, i, over time would tell us whether industry k is becoming more or less concentrated over time at different locations. In a practical applications looking at many industries and cities over time or across countries, the issue concerns how to produce summary measures to describe either how overall concentration varies across industries or how one city's specialization compares with another's. Another issue concerns how to factor in the different forces that cause specialization or concentration phenomena. The literature uses a variety of approaches. We start by looking at urban specialization.

1.2.1 Urban Specialization

Evidence on countries such as Brazil, U.S.A., Korea, and India (Henderson 1988, and Lee 1997) indicate that cities are relatively specialized. The traditional urban specialization literature going back to Bergsman, Greenston and Healy (1972) uses cluster analysis to group cities into categories based on similarity of production patterns -- correlations (or minimum distances) in the shares of different industries in local employment, s_{ik}. Cluster analysis is an "art form" in the sense that there is no optimal set of clusters, and it is up to the researcher to define how fine or how broad the clusters should be and there are a variety of clustering algorithms.

Using 1990 data for the U.S.A., Black and Henderson (2003) group 317 metro areas into 55 clusters, "defining" 55 city types based on patterns of specialization for 80 2-digit industries. They define textile, primary metals, machinery, electronics, oil and gas, transport equipment, health services, insurance, entertainment, diversified market center, and so on type cities, where anywhere from 5-33% of local employment is typically found in just one industry. They show that production patterns across the types are statistically different and that average cities and educational levels by type differ significantly across many of the types. Specialization especially among smaller cities tends to be absolute. At a 3-digit level many cities have absolutely zero employment in a variety of categories. So in the 1992 Census of Manufactures for major industries like computers, electronic components, aircraft, instruments, metal working machinery, special machinery, construction machinery, and refrigeration machinery and equipment, respectively, of 317 metro areas 40%, 17%, 42%, 15%, 77%, 15%, 14% and 24% have absolutely zero employment in these industries.

Kim (1995) in looking at the USA examines how patterns of specialization have changed over time, by comparing for pairs (i, j) of locations $\sum_k |s_{ik} - s_{jk}|$ and by estimating spatial Gini's for industry concentration. He finds that states are substantially less specialized in 1987 than in 1860, but that localization, or concentration has increased over time. For Korea, as part of the deconcentration process noted earlier, Henderson, Lee, and Lee (2001) find that from 1983 to 1993, city specialization as measured by a normalized Hirschman-Herfindahl index

$$g_j = \sum_k (s_{jk} - s_j)^2 \qquad (4)$$

rises in manufacturing, while a provincial level index declines. Cities become more specialized and provinces less so. Clearly the geographic unit of analysis matters, as do the concepts. City specialization as envisioned in the models presented below is consistent with regional diversity, when large regions are composed of many cities of different types.

Henderson (1997) for the USA and Lee (1997) for Korea show that the g_j index of specialization in manufacturing declines with metro area size. Smaller cities are much more specialized than larger cities in their manufacturing production. More generally, Kolko (1999) demonstrates that larger cities are more service oriented and smaller ones more manufacturing oriented. For six size categories (over 2.5 million, 1 - 2.5 million, ... < .25 million, non-metro counties) Kolko shows that the ratio of manufacturing to business service activity rises from .68 to 2.7 as size declines, where manufacturing and business services account for 35% of local private employment. The other 65% of local employment is in "non-traded" activity whose shares don't vary across cities – consumer services, retail, wholesale, construction, utilities.

1.2.2 Geographic Concentration

What about concentration of industry -- the extent to which a particular industry is found in a few versus many locations? In an extremely important paper Ellison and Glaeser (1999) model the problem using USA data, to determine the extent of clustering of plants within an industry due to either industry-specific natural advantages (e.g., access to raw materials) or spillovers among plants. Plants locate across space so as to maximize profits and profits depend on area specific natural advantage, spillovers, and an i.i.d. drawing from Weibul distribution. The idea is to explain the joint importance of spillovers and natural advantage in geographic concentration.

Geographic concentration for industry j is $G_j = \sum_i (s_{ji} - x_i)^2$, where s_{ji} is the share of industry j in employment in location i and x_i is location i's share in total national employment (to standardize for location size). Where $0 \leq \gamma^{na} \leq 1$ represents the importance of natural advantage (where the variance in relative profitability of a location is proportional to γ^{na}) and γ^S represents the fraction of pairs of firms in an industry between which a spillover exists, under their assumptions, Ellison and Glaeser show that

$$E\left[G_j\right] = (1 - \sum_i x_i^2)\,(\gamma_j + (1-\gamma_j)\,H_j)$$
$$\gamma \equiv \gamma^{na} + \gamma^s - \gamma^s\gamma^{na} \qquad (5)$$

where H_j is the standard Hirschman-Herfindahl index of plant industrial concentration in industry j. So $E\left[G_j\right]$ adjusts γ_j for variations in location size $(1 - \sum x_i^2)$ and industry concentration H. Using (5) and estimates of G_j, H_j, and $(1 - \sum x_i^2)$, the empirical part of their paper calculates γ_j for all 3- or 4-digit manufacturing industries across states and countries. They show for 4-digit industries that $G > (1 - \sum x_i^2)H$ in 446 of 459 industries, where $G \leq (1 - \sum x^2)H$ only if $\gamma \leq 0$. That is, almost all industries display some degree of spatial concentration due to either natural advantage or spillovers. Second they argue that 25% of industries are highly concentrated $(\gamma > .05)$ and 43% are not highly concentrated $(\gamma < .02)$. In a later article, Ellison and Glaeser (1999) argue that, based on econometric results relating location choices to natural advantage measures, 10-20% of γ in eq. (5) is accounted for by natural advantage. The rest is due to intra-industry spillovers, a rather critical finding in urban analysis indicating the importance of understanding the nature of scale externalities.

In an important working paper, Duranton and Overman (2004), look at geographic concentration using British data. Rather than model the underlying stochastic process of industrial location under

specific assumptions to yield a specific index, Duranton and Overman take a non-parametric approach, where they also focus on how to test statistically whether industries are significantly concentrated. They calculate the distribution of all pair-wise distances between plants in an industry. Distributions shifted to the left have a greater concentration of short pair-wise distances and are more spatially concentrated. The authors have the advantage of knowing "exact" plant locations (basically within a city block or so), rather than having to rely on, say, county locations which in the US can cover vast distances. They develop a framework to test observed industry distributions against the "counterfactual" of what distributions would look like if firms choose locations randomly, given (a) the set of locations in the UK for industrial plants is limited, (b) bilateral distances between all possible points are not independent, and (c) industry sizes or numbers of plants differ. The framework involves repeated sampling for an industry without replacement from the set of national industrial sites with the sample size equal to industry size. Following that procedure, they construct 95% confidence intervals to test if observed distributions depart from randomness. Compared to Ellison-Glaeser

In practical applications their approach captures a nuanced aspect of spatial clustering. For relatively concentrated industries, the Ellison and Glaeser index is typically dominated by the county with the highest share (given squared shares in the index), telling us the extent to which an industry is concentrated in just one place. The Duranton-Overman approach tells us more generally about spatial clustering over the whole country. So in Ellison and Glaeser, an industry which has a high concentration in one county but is otherwise very dispersed across the 3000 USA counties may look more concentrated than an industry which is concentrated in, say 3-4 nearby counties, with little representation elsewhere. But the latter would be well represented in Duranton and Overman.

1.2.3 Geography

A variety of recent studies have examined the role of geography, primarily natural features, in the spatial configuration of production and growth of cities. Rappaport and Sacks (2001) building on Sacks' general geography program herald the role of coastline location in the U.S.A., as a factor promoting city growth. In a related study, Beeson, DeJong and Troeskan (2001) look at USA counties from 1840-1990. They show that iron deposits, other mineral deposits, river location, ocean location, river confluence, heating degree days, cooling degree days, mountain location, and precipitation all affect 1840 county population significantly. However for 1840-1990 growth in county population, only ocean location, mountain location, precipitation, and river confluence matter, controlling for 1840 population. That is, first nature items strongly affected 1840 and hence indirectly 1990 populations; but growth from 1840-1990 is independent of many first nature influences. Ocean location as Sacks' suggests has persistent growth effects.

Both these studies ignore the geography of markets and the role of neighbors in influencing city evolution. Dobkins and Ioannides (2001) show that growth of neighboring cities influence own city growth and cities with neighbors are generally larger than isolated cities. Black and Henderson (2003) put neighbor and geographic effects together. They calculate normalized market potential variables (sum of distance discounted populations of all other counties in each decade, normalized across decades). They find climate and coast affect relative city growth rates; but market potential has big effects as well, although they are non-linear. Bigger markets provide more customers, but also more competition, so marginal market potential effects diminish as market potential increases. High market potential helps explain why North-East cities in the USA maintain reasonable growth given it is the most densely populated area from history, despite the hypothesized natural advantages of the West.

1.3 Urbanization in Developing Countries

Urbanization, or the shift of population from rural to urban environments, is typically a transitory process, albeit one that is socially and culturally traumatic. As a country develops it moves from labor-intensive agricultural production to labor being increasingly employed in industry and services. The latter are not land-intensive and are located in cities because of agglomeration economies. Thus urbanization moves populations from traditional rural environments with informal political and economic institutions to the relative anonymity and more formal institutions of urban settings. That in itself requires institutional development within a country. It spatially separates families, particularly intergenerationally as the young migrate to cities and the old stay behind.

Urbanization is a spatial transition process. By upper middle income ranges, countries become "fully" urbanized, in the sense that the percent urbanized levels out at 60-90% of the national population living in cities. The actual percent urbanized with full urbanization varies with geography, the role of modern agriculture in the economy, and national definitions of urban. This idea of a transitory phenomenon is illustrated Figure 2, comparing different regions of the world in 1960 versus 1995. While urbanization increased in all regions of the world over those 35 years, among developed countries there is little change since 1975. By 1995 Soviet bloc and Latin American countries had almost converged to developed country urbanization levels. Only sub-Saharan African and Asian countries still face substantial urbanization in the future. Although urbanization is transitory, given the total spatial transformation and accompanying institutional and social transformation involved, as a policy issue urbanization is very important to developing countries. Here we review some basic facts and issues about the process.

1.3.1 Issues Concerning Overall Urbanization

As noted above, urbanization is the consequence of changes in national output composition from rural agriculture to urbanized modern manufacturing and service production. As such, Renaud (1981) makes the basic point that government policies bias, or influence urbanization through their effect on national sectoral composition. So policies affecting the terms of trade between agriculture and modern industry or between traditional small town industries (textiles, food processing) and high tech large city industries affect the rural-urban or small-big city allocation of population. Such policies include tariffs, and price controls and subsidies. The idea that government policies affect urbanization primarily through their effect on sector composition is a key point of empirical studies of urbanization by Fay and Opal (1999) and Davis and Henderson (2003). These studies show that, indeed, urbanization which occurs in the early and middle stages of development is determined largely by changes in national economic sector composition and in technology, and government policies tend to affect urbanization only indirectly through their effect on sector composition.

Urbanization promotes benefits from agglomeration such as localized information and knowledge spillovers and thus efficient urbanization promotes economic growth. Writers such as Gallup, Sacks and Mellinger (1999) go further to suggest that urbanization may "cause' economic growth, rather than just emerge as part of the growth process. The limited evidence so far suggests urbanization doesn't cause growth per se. Henderson (2003) finds no econometric evidence linking the extent of urbanization to either economic or productivity growth or levels. That is if a country were to enact policies to encourage urbanization per se, typically that wouldn't improve growth.

Finally on urbanization, there is an informal notion (Mills and Becker 1986 and World Bank 2000) that the transitory urbanization process follows the same stages as population growth (the "demographic" transition between falling death rates and falling fertility rates) – an S-shaped relationship where urban population growth is slow at low levels of development, then there is a period of rapid acceleration in intermediate stages, followed by a slowing of growth. However the data suggest otherwise at least over the last 35 years. Figure 3 illustrates after parceling out the effect of national population, or country size, based on pooled country data every 5 years from 1965-1995. In Figure 3 the log of national urban population is an increasing concave function of the log of income per capita, indicating the *growth* rate of urban population is a concave increasing function of income levels (Davis and Henderson, 2003).

1.3.2 The Form of Urbanization: The Degree of Spatial Concentration

In 1965, Williamson published an innovative paper based on cross-section analysis of 24 countries in which he argued that national economic development is characterized by an initial phase of internal regional divergence, followed by a phase of later convergence. That is, a few regions initially experience accelerated growth relative to other (peripheral) regions, but later the peripheral regions start

to catch up. Barro and Sala-i-Martin (1991 and 1992) present extensive evidence on this for the USA, Western Europe, and Japan, by examining the evolution of inter-regional differences in per capita incomes. While inter-regional out-migration from poorer regions plays a role in catch-up, it may not be critical. For Japan, the authors argue that later convergence of backward regions occurred through improved productivity in backward regions.

The urban version of this divergence-convergence phenomenon looks at urban primacy. Following Ades and Glaeser (1995), conceptually the urban world is collapsed into two regions -- the primate city versus the rest of the country, or at least the urban portion thereof. The basic question concerns to what extent urbanization is concentrated, or confined to one (or a few) major metro areas, relative to being spread more evenly across a variety of cities. Primacy is commonly measured by the ratio of the population of the largest metro area to all urban population in the country (Ades and Glaeser 1995, Junius 1999, and Davis and Henderson 2003). A more comprehensive measure might use either a spatial Gini or a Hirschman-Herfindal index [HHI] from the industrial organization literature.

Corresponding to Williamson's hypothesis, these papers find an inverted U-shape relationship, where relative urban concentration first increases, peaks, and then declines with economic development. Despite different concentration measures and methods, Wheaton and Shishido (1981) examining a HHI using cross-section non-linear OLS and Davis and Henderson (2003) examining primacy using panel data methods and IV estimation find that concentration rises, peaks in the $2000-4000 range (1985 PPP dollars), and then declines. As Figure 4 illustrates, without conditioning on other variables affecting primacy, the inverted $U-$ relationship of primacy against income is noisy and only apparent in the raw data in earlier time periods (cf. 1965-75 in part (a) with 1985-95 in part (b)).

Lee (1997) explores the relationship between changes in urban concentration and industrial transformation for Korea. The idea is that manufacturing is also first very concentrated in primate cities at early stages of development and then decentralizes to such an extent that at the other end of economic development it is relatively more concentrated in rural areas, as in the USA today as noted earlier. Seoul's urban primacy peaked around 1970 and while Seoul's absolute population has continued to grow, its share has declined steadily. At the urban primacy peak in 1970, Seoul had a dominant share of national manufacturing although the other major metro areas, Pusan and Taegu, also had large shares. During the next 10-15 years, manufacturing suburbanized from Seoul to satellite cities in the rest of Kyonggi province (its immediate hinterland), as well as to satellite cities surrounding Pusan and Taegu. Such suburbanization of manufacturing has been documented also for Thailand (Lee, 1988), Colombia (Lee, 1989), and Indonesia (Henderson, Kuncoro and Nasution, 1996). But the key development following the early 1980's in Korea is the spread of manufacturing from the three major metro areas (Seoul, Pusan, and Taegu) and their satellites to rural areas and other cities. The share of rural areas and other cities in

manufacturing rose from 26% in 1983 to 42% in 1992, in a time period when national manufacturing employment is fairly stagnant and rural areas and other cities actually continue to experience modest absolute population losses. That is, manufacturing deconcentrated both relatively and absolutely to hinterland regions. This deconcentration coincided with economic liberalization, enormous and widespread investment in inter-regional transport and infrastructure investment, and fiscal decentralization (Henderson, Lee, and Lee, 2001) and is consistent with core-periphery reversal in the new economic geography literature discussed later.

Given the urban primacy relationships, the immediate issue is the "so-what" question. How is urban concentration important to growth? For example, is there an optimal degree of urban primacy with each level of development, where significant deviations from this level detract from growth? Conceptually there should be an optimal degree of primacy, where optimal primacy involves a trade-off of the benefits of increasing primacy-- enhanced local scale economies contributing to productivity growth-- against the costs -- more resources diverted away from productive and innovative activities to shoring up the quality of life in congested primate cities. In the first econometric examination of this so-what question, Henderson (2003), using panel data and IV estimation for 1960-1990, finds that there is an optimal degree of primacy at each level of development that declines as development proceeds. Optimal primacy is the level that maximizes national productivity growth. Initial high relative agglomeration is important at low levels of development when countries have low knowledge accumulation, are importing technology, and have limited capital to invest in widespread hinterland development. However the desirability of high relative agglomeration declines with development. Error bands about optimal primacy numbers are quite tight. Second, large deviations from optimal primacy strongly affect productivity growth. An 33% increase or decrease in primacy from a typical best level of .3 reduces productivity growth by 3% over five years. There is some tendency internationally to excessive primacy, with the usual suspects such as Argentina, Chile, Peru, Thailand, Mexico, and Algeria having extremely high primacy.

Why would countries significantly deviate from desired levels of concentration? There is a considerable literature of government policies and institutions in fostering excessive concentration. In Ades and Glaeser (1995), the basic idea is that national policy makers favor the national capital (or other seat of political elites such as São Paulo in Brazil) for reasons of personal gain. For example, direct restraints on trade for hinterland cities such as inability to access capital markets or to get export or import licenses favor firms in the national capital. Policy makers and bureaucrats may gain as shareholders in such firms, or they may gain rents from those seeking licenses or other exemptions to trade restraints (see Henderson and Kuncoro, 1996, on Indonesia). Indirect trade protection for the primate city can also involve under-investment in hinterland transport and communications infrastructure.

Whether as true beliefs or as a justification to cover rent-seeking behavior, policy makers in different countries often articulate a view that large cities are more productive and thus should be the site for government-owned heavy industry (e.g., São Paulo or, Beijing-Tianjin historically). Later we will point out that it may be that output per worker in heavy industries is higher in the productive external environment of large metro areas. It just isn't high enough to cover the higher opportunity costs of land and labor in those cities, which is one reason why those state-owned heavy industries lose money in such cities.

Favoritism of a primate city creates a non-level playing field in competition across cities. The favored city draws in migrants and firms from hinterland areas, creating an extremely congested high cost-of-living metro area. Local city planners can try to resist the migration response to primate city favoritism by, for example, refusing to provide legal housing development for immigrants or to provide basic public services in immigrant neighborhoods. Hence the development of squatter settlements, bustees, kampongs and so on. But still, favored cities tend to draw in enormous populations.

Is there econometric evidence indicating that politics plays a role in increasing sizes of primate cities? Ades and Glaeser (1995) based on cross-section analyses find that if the primate city in a country is the national capital it is 45% larger. If the country is a dictatorship, or at the extreme of non-democracy, the primate city is 40-45% larger. The idea is that representative democracy gives a political voice to the hinterland regions limiting the ability of the capital city to favor itself. Apart from representative democracy, fiscal decentralization helps to level the playing field across cities, by giving political autonomy for hinterland cities to compete with the primate city.

Davis and Henderson (2003) explore these ideas further, examining in a panel context the impact upon primacy of democratization and fiscal decentralization from 1960-1995. Using a panel approach with IV estimation, they find smaller effects than Ades and Glaeser but still highly significant ones. Examining both democratization and fiscal decentralization together, they find moving from the extreme of least to most democratic form of government reduces primacy by 8% and from the extreme of most to least centralized government reduces primacy by 5%. Primate cities which are national capitals are 20% larger and primate cities in planned economies with migration restrictions are 18% smaller. Finally they find transport infrastructure investment in hinterlands which opens up international markets to hinterland cities reduces primacy, as the core-periphery models of the new economic geography tend to predict. A one-standard deviation increase in roads per sq. kilometer of national land area or in navigable inland waterways per sq. kilometer each reduce primacy by 10%.

2. Cities and Growth

To establish the links between cities, growth, urbanization, urban concentration and policy, we look at models in which cities are a defined unit, endogenous in number and size. These are systems of cities models which date to Henderson (1974), with a variety of substantial contributors to further development (Hochman 1977, Kanemoto 1980, Henderson and Ioannides 1981, Abdel-Rahman and Fujita 1990, Helsley and Strange 1990, Duranton and Puga 2000, and Rossi-Hansberg and Wright 2004, to name a few). Here I outline the model in Black and Henderson (1999a) which is an endogenous growth model of cities, examining the growth-urban connection. The analysis is broken into several parts. The first reviews the traditional static model, focused on city formation and the determination of the sizes, numbers, and industrial composition of cities in an economy at a point in time. A thorough review of static models is in Abdel-Rahman and Anas (2004), so our treatment focuses on what we need to analyze growth and later urbanization and development. We then turn to the growth part, focusing first on steady-state growth and a variety of extensions covering stochastic processes and analysis of functional specialization. Section 3 turns to rural-urban transformation, or urbanization under economic development. That section also discusses issues of city debt finance and land market institutions.

2.1 The Systems of Cities at a Point in Time

Consider a large economy composed of two types of cities, where there are many cities of each type and each type is specialized in the production of a specific type of traded good. We will show why (when) there is specialization momentarily; the generalization to many types of goods and cities is straightforward. To simplify the growth story, each firm is composed of a single worker. In a city type 1, in any period, the output of firm i in a type 1 city is

$$X_{1i} = D_1 \ (n_1^{\delta_1} h_1^{\psi_1}) \ h_{1i}^{\phi_1}, \ 0 < \delta_1 < \frac{1}{2} \qquad (6)$$

h_{1i} is the human capital of the worker and is his input in the production process. A worker-firm is subject to two local externalities. First is own industry localization economies, the level of which depends on the total number of worker-firms, n_1, in this representative type 1 city. There is a large literature on micro-foundations of localization economies, with an excellent analysis and review in Duranton and Puga (2004). While the concepts are discussed in Marshall (1890), the formal literature dates to Fujita and Ogawa (1982) who model micro-foundations as exogenous information spillovers that enhance productivity but decay with spatial distance between plants. Such spillovers can be made endogenous (Kim, 1988) with the volume of costly "contacts" being a firm choice variable. But the modern literature

on micro-foundations as reviewed in Duranton and Puga moves on to try to model why contacts matter, rather than just assuming they matter.

In this section spatial decay is all or nothing – no decay within the city's; 100% across cities. As such in (6), n_1 could represent the total volume of local spillover communications, where δ_1 is the elasticity of firm output with respect to n_1. The restriction $\delta_1 < 1/2$ ensures a unique solution in an economy composed of many type 1 cities. A larger δ_1 results in all X_1 production crowding into one city. Note the production process ignores land, collapsing the central business district [CBD] to a point. There is a recent literature building upon Fujita and Ogawa where firm density is endogenous in a spatial CDB with information spillover decay. There market equilibrium density is non-optimal because firms in making location decisions don't recognize that choices leading to greater densities would enhance information spillovers (Lucas and Rossi-Hansberg, 2002 and Rossi-Hansberg, 2004). The issue of central city design and zoned density is an important one in the design of cities in developing countries. But this review focuses on other issues.

The second externality in (6) derives from h_1, the average level of human capital in the city, which represents local knowledge spillovers. $h_1^{\psi_1}$ could be thought of as the richness of information spillovers $n_1^{\delta_1}$, so that knowledge enhances (multiplies) local information spillovers, or gives better information. Alternatively it could just represent the level of local technology, which increases as average education increases locally.

Given this simple formulation the wage of worker i in city type 1 is

$$W_{1i} = X_{1i} \tag{7}$$

In an economy of identical individual workers in type 1 cities, individuals will all have the same human capital level (either exogenously in a static context, or endogenously in a growth context). Thus total city output is

$$X_1 = D_1 h_1^{\phi_1 + \psi_1} n_1^{1+\delta_1} \tag{8}$$

2.1.1 Equilibrium City Sizes

Equations (6) and (8) embody the scale benefits of increases in local employment, where output per worker is an increasing function of local own industry scale. Determinant city sizes arise because of scale diseconomies in city living, including per capita infrastructure costs, pollution, accidents, crime, and

commuting costs. In Henderson (1974) those are captured in a general cost of housing function, but most urban models consider an explicit internal spatial structure of cities. As noted all production occurs at a point, the CBD. Surrounding the CBD in equilibrium in local land markets is a circle of residents each on a lot of unit size. People commute back and forth at a constant cost per unit (return) distance of τ. That cost can be from working time, or here an out-of-pocket cost paid in units of X_1. Equilibrium in the land market is characterized by a linear rent gradient, declining from the center to zero at the city edge where rents (in agriculture) are normalized to zero. Standard analysis dating to Mohring (1961) gives us expressions for total city commuting and rents, in terms of city population where[4]

$$\text{total commuting costs} = bn_1^{3/2} \tag{9}$$

$$\text{total land rents} = 1/2\ bn_1^{3/2} \tag{10}$$

$$\text{b} \equiv 2/3\ \pi^{-\frac{1}{2}}\tau.$$

Equation (9) is the critical resource costs, where marginal commuting costs are increasing in city population. Rents are income to, potentially, a city developer.

How do cities form and how are sizes determined? We start with a specific mechanism and discuss how it generalizes below, and what happens if such a mechanism isn't present. There is an unexhausted supply of identical city sites in the economy, each owned by a land developer in a nationally competitive urban land development market. A developer for an occupied city collects local land rents, specifies city population (but there is free migration in equilibrium), and offers any inducements to firms or people to locate in that city, in competition with other cities in order to maximize profits. Population is freely mobile.

The land developer maximizes

[4] An equilibrium in residential markets requires all residents (living on equalize size lots) to spend the same amount on rent, $R(u)$, plus commuting costs, τu, for any distance u from the CBD. Any consumer then has the same amount left over to invest or spend on all other goods. At the city edge at a radius of u, rent plus commuting costs are τu_1 since $R(u_1) = 0$; elsewhere they are $R(u) + \tau u$. Equating those at the city edge with those amounts elsewhere yields the rent gradient $R(u) = \tau(u_1 - u)$. From this, we calculate total rents in the city to be $\int_0^{u_1} 2\pi u R(u) du$ (given lot sizes of one so that each "ring" $2\pi u du$ contains that many residents) or $1/3\pi\tau u_1$. Total commuting costs are $\int_0^{u_1} 2\pi u\ (\tau u)\ du = 2/3\pi r u_1^3$. Given a city population of n and lot sizes of one, $n_1 = \tau u_1^2$ or $u_1 = \pi^{-\frac{1}{2}} n^{\frac{1}{2}}$. Substitution gives us eqs. (9) and (10).

$$\max_{n_1,\ T_1} \ profit_1 = 1/2\ bn_1^{3/2} - T_1 n_1$$
$$\text{subject to } W_1 + T_1 - 3/2\ bn_1^{\frac{1}{2}} = I_1 \tag{11}$$

where T_1 is the per firm subsidy (e.g., in practice, in a model with local public goods, a tax exemption). I_1 is the real income per worker available in equilibrium in national labor markets under free mobility, which a single developer takes as given. In the constraint, I_1 equals wages in (7), plus the subsidy, less per worker rents plus commuting costs paid from (9) and (10). Maximizing with respect to T_1 and n_1 and imposing perfect competition in national land markets so $profit_1 = 0$ ex post, yields

$$T_1 = 1/2\ bn_1^{\frac{1}{2}} \tag{12}$$

$$n^* = (\delta_1 2b^{-1} D_1)^{2/(1-2\delta_1)} h_1^{2\varepsilon_1} \tag{13}$$

$$\varepsilon_1 \equiv \frac{\phi_1 + \psi_1}{1 - 2\delta_1} \tag{14}$$

This solution has a variety of properties heralded in the urban literature. First it reflects the Henry George Theorem (Flatters, Henderson, and Mieszkowski 1974, Stiglitz 1977), where the transfer per worker/firm exactly equals the gap $(\delta_1 W_1)$ between social and private marginal of labor to the city, and that subsidy which prices externalities is exactly financed out of collected land rents at efficient city size. That is, total land rents cover the cost of subsidies needed to price externalities, as well as the costs of local public goods in a model where good goods are added in. Second the efficient size in (13) is the point where real income, I_1, peaks, as an inverted $U-$shape function of city size, as we will illustrate later in Figure 5. If $\delta_1 < 1/2$, we can show that I_1 is a single-peaked function of n_1, so n_1^* is the unique efficient size. If $\delta_1 > 1/2$, in essence there will only be one type 1 city in the economy, because net scale economies are unbounded. Given n_1^* is the size where I_1 peaks, n_1^* is a free mobility equilibrium -- a worker moving to another city would lower real income in that city and be worse off. Finally city size is increasing in technology improvements: τ declining, δ_1 rising, D_1 rising, or local knowledge accumulation (h_1) rising.

By substituting in the constraint in (11), we can define relationships among real income, wages, and human capital. Substituting in first for T_1 and then n_1 we get

$$I_1 = W_1 - bn^{1/2} = (1-2\delta)\ W_1 = Q_1 h_1^{\varepsilon_1} \tag{15}$$

where Q_1 is a parameter cluster. Note real income is wages deflated by urban living costs; and that real income rises with human capital.

Institutions and City Size. I have specified the equilibrium in national land markets, given competitive developers. Helsley and Strange (1990) put this in proper context, specifying the city development game, determining how many cities will form and what their sizes (n^*) will be. Henderson and Becker (2001) show that the resulting solutions (with multiple factors of production) are (1) Pareto efficient, (2) the only coalition proof equilibria in the economy, (3) unique under appropriate parameters, and (4) free mobility ones where the developer specified populations are self-enforcing. They also show that, under appropriate conditions, such outcomes arise (1) in a self-organized economy with no developers where city governments can exclude residents ("no-growth" restrictions) to maximize the welfare of the representative local voter, (2) in a growing economy where developers form new cities and old cities are governed by local governments. Note for developing countries the key ingredients: either national land markets must be competitive with developers free to form new cities or atomistic settlements can arise freely and local autonomous governments can limit their populations as they grow.

Absent such institutions, cities only form through "self-organization". In the model here with perfect mobility of resources, the result is potentially enormously oversized cities (Henderson 1974, Henderson and Becker 2001). Nash equilibrium city size in atomistic worker migration decisions lies between efficient size, n^*, and a limit size to the right, $n_{\max}$, where city size is so large with such enormous diseconomies that the population is indifferent between being in a rural settlement of size 1 (the size of a community formed by a defecting migrant) and $n_{\max}$ That is, given an inverted-U shape to real income I_1, self organization has cities at the right of the peak n^*, potentially at $n_{\max}$ where where $I_1\ (n=1) = I_1\ (n=n_{\max})$. The problem is one of co-ordination failure.

Consider a large economy with growing population, where, in size, all cities are at or just beyond n^*. Timely formation of the next city to accommodate this population growth requires en mass movement of population from existing cities into a new city of size, n^*. Without co-ordination in the form of developers or city governments, no such en mass movement is possible, so people wait to migrate from existing cities to a new city until existing cities have all grown to $n_{\max}$, where it pays individual

migrants to exit to a cities to set up their own tiny "city". Of course at that "bifurcation point" (Krugman, 1991a), in equilibrium these milling migrants coalesce into 1 or more new cities of size greater than or equal to n^*, at which point, again, all then existing cities start to grow again with national population growth until they too hit the bifurcation point $n_{\max}$. This dismal process is what faces countries where local autonomy and national markets are poorly functioning, so that there are no market or institutional mechanisms to co-ordinate en mass movements of people. However the process we have outlined involving population swings across cities and potentially enormously over-populated cities may not be consistent with the data. In Section 3 we will outline a model with immobile capital, where self-organization can involve "commitment" given irreversibility of investment decisions. In that context outcomes, while still inefficient, are not so dismal.

2.1.2 Other City Types

In Black and Henderson, X_1 city type 1 is an input into production of the single final good in the economy, X_2 (from which, hence in a growth context human capital is also "produced"). In many models all outputs of specialized city types are final consumption goods. But here we follow Black and Henderson, without loss of generality. X_2 is produced in type 2 cities where the output for worker/firm j is correspondingly

$$X_{2j} = D_2 \ (n_2^{\delta_2} h_2^{\psi_2}) \ h_{2j}^{\theta_2} \ X_1^{1-\alpha} \tag{16}$$

As in type 1 cities, per worker output is subject to own industry local scale externalities and to local knowledge spillovers. However now there is an intermediate input, X_{1j}, which is the numeraire good with X_{2j} priced at P in national markets. The analysis of city sizes and formation for type 2 cities proceeds as for city type 1, with corresponding expressions, other than the addition of an expression for P in n_2^* and I_2 and a restriction for an inverted $U-$shape to I_2 that $\delta_2 < \alpha/2$.

In a static context the model is closed by utilizing the national full employment constraint

$$m_1 n_1 + m_2 n_2 = N \tag{17}$$

where m_1 and m_2 are the numbers of each type of city and N is national population. The second equation (to solve the 3 unknowns P, m_1 and m_2) equates real incomes as in equation (15) across cities $(I_1 = I_2)$, where individual workers move across cities to equalize real incomes. Finally there is an

equation where national demand equals supply in either the market. That is, the supply, $m_1 X_1$, equals the demand for X_1 as an intermediate input, $m_2 n_2 x_{1,}$ and for producing commuting costs $(m_1\ (bn_1^{3/2}) + m_2\ (bn_2^{3/2}))$ from eq. (9). In this specific model, the solution yields values of m_1, m_2 and *P* that are functions of parameters and h_1 and h_2. In a static context of identical workers, one would impose $h = h_1 = h_2$. We will discuss momentarily the solution for h_1 and h_2 and the model in the growth context. Later in section 3, we will detail solutions for prices and numbers of cities in a simpler but related two sector model. Here given log-linear production functions and a single final consumption good, as Black and Henderson show, X_1 and m_1 / m_2 will be constant over time, independent of *h.*

In the static context where, labor mobility requires $I_1 = I_2$, in the larger type of city, say type 1, commuting and land rent costs will be higher. Thus, if real incomes are equalized, from (15) , $W_1 > W_2$ as a compensating differential for higher living costs. Firms in type 1 cities are willing to pay higher wages because type 1 cities offer them greater scale benefits. Empirical evidence shows as cities move from a small size (say, 50,000) to very large metro areas, the cost-of-living typically doubles (Thomas 1978, Henderson, 2002), explaining the fact that nominal wages also double.

Another issue discussed at length in section 1.3 is that policy makers may favor large cities because they view them as "more productive". Indeed for an industry found in smaller towns, it may be that the externalities they face in equations (6) or (16) may be higher in a larger city. However that doesn't mean they locate there. Although externalities may be higher, in order for them to locate there, it must be sufficiently relatively higher to afford the higher wage and land rents, compared to a smaller city. If not, their profit maximizing or cost minimizing location is the smaller city.

Specialization. This analysis presumes cities specialize in production. That is an equilibrium outcome under a variety of conditions. In the model described so far, there are no costs of inter-city trade: no costs of shipping X_1 as inputs to X_2 type cities and shipping X_2 back as retail goods in X_1 type cities. All transport costs are internal to the city, given the relative greater importance of commuting costs in modern economies. Given that and given scale economies are internal to the industry, any specialized city (formed by a developer) out-competes any mixed city. The heuristic argument is simple. Consider any mixed city with $\tilde{n}_1$ and $\tilde{n}_2$ workers in industry 1 and 2. Split that city into two specialized cities, one with just $\tilde{n}_1$ people and the other with just $\tilde{n}_2$. Scale economies are undiminished ($\tilde{n}_1^{\delta_1}$ and $\tilde{n}_1^{\delta_2}$ in both cases in industries 1 and 2 respectively) but per worker commuting costs are lower in the specialized cities compared to the old larger mixed cities, so real incomes are higher in each specialized city compared to the old city.

Having own industry, or localization economies is a sufficient but not necessary condition for specialization. Industries can instead all have "urbanization" economies where scale depends on total local employment. However if the degree of urbanization economies differs across industries then each industry has a different efficient local scale and is better off in a different size specialized city than any mixed city. Mixed cities occur more in situations where each good has localization economies enhanced by separate spillovers from the other industry or sharing of some common public infrastructure (Abdel-Rahman 2000).

A basic problem in these urban models is the lack of nuance on transport costs. Either transport costs of goods across cities is zero ($X_1$and X_2) or infinite (housing, and potentially other non-tradables). A recent innovation is to have generalized transport costs (without a specific geography) where the cost of transporting a unit of X_1 to an X_2city is t_1 and the cost of shipping X_2 back to an X_1 city is t_2, an innovation due to Abdel-Rahman (1996) in a model similar to the one used here (one intermediate and one final good) and then generalized by Xiong (1998) and Anas and Xiong (1999). Now whether there are specialized as opposed to diversified cities depends on the level of t_1 and t_2. At appropriate points as t_1 or t_2 or both rise from zero, X_1 and X_2 will collocate (in developer run cities). More generally with a spectrum of, say, final products, we would expect that some products with low enough t 's will always be produced in specialized cities, some high enough t 's will be in all cities, and some with middle range t 's will be produced in some cities (ones with bigger markets) but not others (with smaller markets). No one has yet simulated this more complex outcome.

2.1.3 Replicability and National Policy

At the national level in a large economy with many cities, at the limit, there are constant returns to scale, or replicability. If national population doubles, the numbers of cities of each type and national output of each good simply doubles, with individual city sizes, relative prices and real incomes unchanged .[5] With two goods and two factors, basic international trade theorems (Rybczynski, factor price equalization, and Stolper-Samuelson) hold (Hochman 1977, Henderson 1988). This gives an urban flavor to national policies (Renaud 1981, Henderson, 1988). For example trade protection policies favoring industry X_1 produced in relatively large size cities over industry X_2 produced in smaller type cities will alter national output composition towards X_1 production and increase the number of large relative to small cities. National urban concentration will rise. Similarly subsidizing an input such as

[5] Here with h_1 and h_2 yet to be solved we would need to double the numbers of people with h_1 and h_2 respectively. Below we will see the solution with growth to h_1 and h_2 is national scale invariant.

capital for a high tech product, X_1, again, say, produced in a larger type of city will cause the numbers of that type of city to increase, raising urban concentration.

2.2 Growth in a System of Cities

Black and Henderson (1999a) specify a dynastic growth model where dynastic families grow in numbers at rate g over time starting from size 1. If c is per person family consumption, the objective function is $\int_0^{\infty} \frac{c(t)^{1-\sigma}-1}{1-\sigma} e^{-(\rho-g)t} dt$ where $\rho(>g)$ is the discount rate. Dynasties can splinter (as long as they share their capital stock on an equal per capita basis) and the problem can be put in an overlapping generations context with equivalent results (Black, 2000), under a Galor and Zeira (1993) "joy of giving" bequest motive.

The only capital in the model is human capital and as such there is no market for it. Intra-family behavior substitutes for a capital market. Specifically families allocate their total stock of human capital (H) and members across cities, where Z proportion of family members go to type 1 cities (taking $Z\ h_1 e^{gt}$ of the H with them) and $(1-Z)$ go to type 2 cities taking $(1-Z)\ \mathrm{h}_2 e^{gt}$ with them). Additions to the family stock come from the equation of motion where the cost of additions, PH, equals family income $Ze^{gt} I_1 + (1-Z)\ e^{gt} I_2)$ less the value of family consumption of X_2, or $Pc\ e^{gt}$. Constraints prohibiting consumption of human capital, non-transferability except to newborns, and non-transferability within families across city types (either directly or indirectly through migration) are non-binding on equilibrium paths.

Families allocate their populations across types of cities, with low human capital types (say h_1) "lending" some of their share $(h = H / e^{gt})$ to high human capital types (say h_2). High human capital types with higher incomes $(I_2 > I_1$ if $h_2 > h_1)$ repay low human capital types so $c_1 = c_2 = c$ (governed by the family matriarch). This in itself is an interesting development story, where rural families diversify migration destinations (including the own rural village) and remittances home are a substantial part of earnings. In Black and Henderson if capital markets operate perfectly for human capital (i.e., we violate the "no slavery" constraint) or capital is physical and capital markets operational, one dynastic family could move entirely to, say, type 1 cities and lend some of their human capital to another dynastic family in type 2 cities. With no capital market, each dynastic family must operate as its own informal capital market and spread itself across cities.

In this context Black and Henderson show that, regardless of scale or point in the growth process, h_1 / h_2 and I_1 / I_2 are fixed ratios, dependent on θ_i eq. (6) and (16). As θ_1 / θ_2 (the relative returns to capital) rises, h_1 / h_2, I_1 / I_2 and also n_1 / n_2 rise. Z and m_1 / m_2 are all fixed ratios of parameters θ_i, δ_i, and α under

equilibrium growth. Equilibrium and optimal growth differ because the private returns to education in a city, θ_i, differ from the social returns, $\theta_i + \psi_i$. But local governments can't intervene successfully to encourage optimal growth. Why? With free migration and "no slavery", if a city invests to increase its citizens' education, a person can take their human capital ("brain drain") and move to another city (be subsidized by another city to immigrate, given that city then need not provide extra education for that worker). This model hazard problem discourages internalization of education externalities.

2.2.1 Growth properties: Cities

From eq. (13), equilibrium (and efficient) city size in type 1 cities is a function of the per person human capital level, h_1, in type 1 cities. After solving out the model (for P), the same will be true of type 2 cities. City sizes grow as h_1 and h_2 grow, where, under equilibrium growth given h_1 / h_2 is a fixed ratio, $\dot{h}_1 / h_1 = \dot{h}_2 / \dot{h}_2$ where a dot represents a time derivative. Then

$$\frac{\dot{n}_2}{n_2} = \frac{\dot{n}_1}{n_1} = 2\varepsilon_1 \frac{\dot{h}}{h} \tag{18}$$

where $\dot{n}_i / n_i$ is the growth rate of efficient sizes n_i^*.

For the number of cities, the issue is whether growth in individual sizes absorbs the national population growth, or more cities are needed. Given

$$\frac{\dot{m}_1}{m_1} = \frac{\dot{m}_2}{m_2} = g - \frac{\dot{n}_i}{n_i} = g - 2\varepsilon_1 \frac{\dot{h}}{h}, \tag{19}$$

the numbers of cities grow if $g > \dot{n}_i / n_i$. Note growth in numbers and sizes of cities is "parallel" by type, so the relative size distribution of cities is constant over time. Parallel growth with a constant relative size distribution of cities as reviewed in Section 1.1 is what is observed in the data. This result generalizes to many types of cities under certain conditions. For example, with the log linear production technologies we assumed and with many varieties of output consumed under unitary price and income elasticities of log-linear preferences, parallel growth results.

2.2.2 Growth properties: Economy

Ruling out explosive or divergent growth, there are two types of growth equilibria. Either the economy converges to a steady state level, or it experiences endogenous steady-state

growth. Convergence to a level occurs if $\varepsilon \equiv \varepsilon_1 \ (1-(\gamma - 2\delta_2)) + \varepsilon_2 \ (\gamma - 2\delta_2) < 1,$ where ε is a weighted average of the individual city type. In that case at the steady-state $\bar{h}$, $\dot{n}_i / n_i = 0$ and $\dot{m}_i / m_i = g$, or only the numbers but not sizes of cities grow just like in exogenous growth (Kanemoto 1980, Henderson and Ioannides, 1981). If $\varepsilon = 1$ then there is steady-state growth, where $\bar{\gamma}^h = \dot{h}/h = \frac{A-\rho}{\sigma}$ (where the transversality condition requires $A > \rho$). In that case $\dot{n}_i / n_i = 2\varepsilon_1^{\left(\frac{A-\rho}{\sigma}\right)}$, or cities grow at a constant rate. and their numbers also increase if $g > 2\varepsilon_1^{\left(\frac{A-\rho}{\sigma}\right)}$. This "knife-edge" result of whether there is endogenous growth or not dependent on the value of ε is not essential. For example in Rossi-Hansberg and Wright (2004) endogenous growth can occur more generally in a context where human capital accumulation involves worker time and the growth rate of human capital is a log-linear function of the fraction of time devoted to human capital accumulation, as opposed to production.

2.3 Extensions

There are three major extensions to the basic systems of cities models. First people may differ in terms of inherent productivity or in terms of endowments. Second, while we have discussed the issue of city specialization versus diversification we haven't developed insights into a more nuanced role of small highly specialized cities versus large diversified metro areas in an economy.

2.3.1 Different Types of Workers

Turning to the first extension, Henderson (1974) has physical capital as a factor of production owned by capitalists who needn't reside in cities. Equilibrium city size reflects a market trade-off between the interests of city workers who have an inverted U-shape to utility as a function of the size of the city they live in and capitalists whose returns to capital rise indefinitely with city size (for the same capital to labor ratio). There is a political economy story, where capitalists collectively in an economy have an incentive to limit the number of cities, thus forcing larger city sizes. Helsley and Strange (1990) have a matching model between the attributes of entrepreneurs and workers and Henderson and Becker (2001) a related two class model. Again the two class model yields a conflict between the city sizes that maximize the welfare of one versus another group, which is resolved in competitive national land development markets.

In a different approach Abdel-Rahman and Wang (1997), Abdel-Rahman (2000) and later Black (2000) look at high and low skill workers who are used in differing proportions in production of different goods. Black has one traded good produced with just low skill labor and a second traded good produced with high skill workers and inputs of a local non-traded good produced with just low skill workers. High skill workers generate production externalities in the form of knowledge spillovers for all traded goods. In Black, urban specialization with all high skill workers (and some low skill workers) concentrated in one

type of city producing the first type of good is efficient; but a separating equilibrium that would sustain this pattern, where low skill workers and low tech production stay in their own type of city (rather than trying to cluster with high tech production) is not always sustainable. Black characterizes conditions under which a separating equilibrium will emerge.

It is important to note that there is a much more developed literature on inequality induced by neighborhood selection, where the characteristics of neighbors affect skill acquisition (e.g. average family background in the classroom affects individual student performance). That leads to segregation of talented or wealthier families by neighborhood (Benabou 1993, Durlauf, 1996) and can help transmit economic status across generations, promoting inter-generational income inequality.

2.3.2 Metro Areas.

Simple indices of urban diversity indicate that smaller cities are very specialized and larger cities highly diversified. So the question is what is the role of large metro areas in an economy and their relationship to smaller cities. Henderson (1988) and Duranton (2004) have a first nature - second nature world, where every city has a first nature economic base and footloose industries cluster in these different first nature cities. In general the largest centers are those attracting the most footloose production to their first nature center. The Duranton paper is discussed in detail in Section 2.3.3. However, it seems that today few metro areas have an economic base of first nature activity. Accordingly recent literature has focused on the role of large metro areas as centers of innovation, headquarters, and business services (Kolko, 1999).

The Dixit-Stiglitz model opened up an avenue to look at large metro areas as having a base of diversified intermediate service inputs, which generate scale-diversity benefits for local final goods producers. That initial idea was developed in Abdel-Rahman and Fujita (1990) and has led to a set of papers focused on the general issue of what activities, under what circumstances are out-sourced. Theory and empirical evidence (Holmes, 1999 and Ono 2000) suggest that as local market scale increases, final producers will in-house less of their service functions. The resulting increased out-sourcing encourages competition and diversity in the local business service market, encouraging further out-sourcing.

In terms of incorporating this into the role of metro areas versus smaller cities, Davis (2000) has a two-region model, a coastal internationally exporting region and an interior natural resource rich region. There are specialized manufacturing activities which, for production and final sale, require business service activities, summarized as headquarters functions. Headquarters purchase local Dixit-Stiglitz intermediate services such as R&D, marketing, financing, exporting, and so on. Headquarters' activity is in port cities in the coastal region. The issue is whether manufacturing activities are also in these ports versus in specialized coastal hinterland cities versus in specialized interior cities. Scale economies in

manufacturing and headquarters activities are different and independent of each other, so that, based on scale considerations, these activities would be in separate specialized cities. However if the costs of interaction (shipping manufactured goods to port and transactions costs of headquarters-production facility communication) between headquarters and manufacturing functions are extremely high, then both manufacturing and headquarters activities can be found together in coastal port cities. Otherwise they will be in separate types of cities. In that case, manufacturing cities will be in coastal hinterlands if costs of headquarters-manufacturing interaction are high relative to shipping natural resources to the coast. However if natural resource shipping costs are relatively high, then manufacturing cities will be found in the interior. Duranton and Puga (2001) have a very similar model of functional specialization, without the regional flavor. If there is specialization, then there are headquarter cities where headquarters outsource local services in diversified large metro areas, while production occurs in specialized manufacturing cities.

In a different paper Duranton and Puga (2000) develop an entirely different and stimulating view of large metro areas. In an economy there are m types of workers who have skills each specific to producing one of m products. Specialized cities have one type of worker producing the standardized product for that type of worker subject to localization economies. Diversified cities have some of all types of workers. Existing firms at any instant die at an exogenously given rate; and, in a steady-state, new firms are their replacement. New firms don't know "their type" -- what types of workers they match best with and hence what final product they would be best off producing. To find their type they need to experiment by trying the different technologies (and hence trying different kinds of workers). New firms have a choice. They can locate in a diversified city with low localization economies in any one sector. In a diversified city they can experiment with a new process each period until they find their ideal process. At that point they relocate to a city specialized in that product, with thus high localization economies for that product. Alternatively new firms can experiment by moving from specialized city to specialized city with high localization economies, but face a relocation cost each time. If relocation costs are high, the advantage during their experimental period is to be in a diversified city. This leads to an urban configuration of experimental diversified metro areas and other cities which are specialized in different standardized manufacturing products.

The Duranton and Puga model captures a key role of large diversified metro areas consistent with the data. They are incubators where new products are born and where new firms learn. Once firms have matured then they typically do relocate to more specialized cities. This also captures the product-life cycle for firms in terms of location patterns. Fujita and Ishii (1994) document the location patterns of Japanese and Korean electronics plants and headquarters. In a spatial hierarchy mega-cities house headquarters activities (out-sourcing business services) and experimental activity. Smaller Japanese or

Korean towns have specialized, more standardized high tech production processes and low tech activity is off-shore.

2.3.3 Stochastic Process and Zipf's Law

Gabaix (1999a, 1999b) argues that if, there is a stochastic process where individual city growth rates follow Gibrat's Law—the growth rate in any period is unrelated to initial size -- then the size distribution the emerges will follow Zipf's Law. Beyond specifying a stochastic process where shocks to productivity or preferences follow a random walk, to get the result in a model where there is an endogenous number of cities of efficient sizes, as opposed to just fixing the number of cities (Gabaix 1999a, and Duranton 2004) requires considerable structure, with a variety of such issues being analyzed in Cordoba (2004). We follow Rossi-Hansberg and Wright (2004) who adapt the model we have presented. In their base case there is only human capital; and technology and preferences are log-linear. They have many final output industries and hence types of specialized cities. They group industries and specialized city types into sets, where within each set industries and city types have the *same* technology but each industry draws its own permanent shock each instant. In terms of the shock they assume that $D_1(t)$ in the equivalent of eq. (6) follows a finite order Markov process. Finally and critically to have Gibrat's Law lead to Zipf's Law, they must impose an arbitrary lower bound on the sizes that cities can fall to (Gabaix (1999a). These assumptions lead to Zipf's Law holding for each set of industries and they show one can aggregate across sets of industries to get Zipf's Law in aggregate. It goes without saying many of the assumptions imposed to get Zipf's Law are very strong, a key point made in Cordoba (2004).

In a recent paper, Duranton (2004) tries to model "micro-foundations" for the stochastic process affecting city sizes and as a result ends up modeling an important overlooked aspect of city evolution. Duranton has "first nature" (immobile given natural resource location) production and "second nature" (mobile, or footloose) production in m cities, where m is given by the number of immobile natural resource products, each needing their own city. So, in contrast to Rossi-Hansberg and Wright the number of cities is fixed; but given that restriction a lot is accomplished. In the paper there are $(n >> m)$ products, in a Grossman-Helpman (1991) product quality ladder model. The latest innovation in each product is produced by the monopolist holding the patent and only this top quality is marketed for any product. Investment in innovation to try to move the next step up in the quality ladder in industry k and get the next patent in k, can also lead to the next step up in a different industry -- i.e., there can be cross-industry innovation. For footloose industries, to partake of a winning innovation occurring for industry k in city i, requires industry k production to locate in city i where the innovator is. Presumably co-location of the inventor and production makes the information needed for the transition to mass production cheaper to exchange (e.g., the workers in the innovative firm take over production). Innovation follows a stochastic process where innovation probabilities depend on R&D expenditures. Industry jumps from city to city

according to where the latest innovation is, and city growth also follow a stochastic process. The resulting stochastic process of city growth and decline results in steady state size distributions that are similar to Zipf's Law. Adding in considerations of urban scale economies in the innovation process helps explain the long right tails in actual city size distributions, as they differ from Zipf's Law.

Duranton's formulation has the nice feature that cities have patterns of production specialization which change over time. This seems to fit the data; and Duranton's paper in fact models the evolution of industry structure of cities. We know from Black and Henderson (1999b) and Ellison and Glaeser (1999) that industries move "rapidly" across cities, with city specialization changing over time for cities. Any city is very slow to gain a high share of any particular industry's production (given there are many possible industries to gain a share from) and is very quick to lose a high share (given many competitor cities).

3. Urbanization and Growth

The previous section examined a fully urbanized economy where all production occurs in cities. City sizes grow with improvements in technology but, absent stochastic elements, individual cities grow in parallel, with the relative numbers of different types of cities and the relative size distribution of cities time invariant. Here we examine a non-steady state world in which an economy has an agriculture sector that is shrinking with economic development and an urban sector that is growing. We briefly review traditional dual sector models and the new economic geography models, both of which examine sectoral transformation, but without cities per se and generally without real economic growth. Then we present an endogenous growth model in which there is sectoral change with cities.

3.1 Two Sector Approaches, Without Cities

Urbanization involves resources shifting from an agricultural to an urban sector. The dual economy models dating back to Lewis (1954) look at sectoral change but are really static models. They focus on the question of urban " bias", or the effect of government policies on the urban-rural divide, and the efficient rural-urban allocation of population at a point in time. These two sector models have an exogenously given "sophisticated" urban sector and a "backward" rural sector (Rannis and Fei 1961, Harris and Todaro 1970, and others as now well exposited in textbooks (e.g., Ray 1998)).

In these models, the marginal product of labor in the urban sector is assumed to exceed that in the rural sector. Arbitrage in terms of labor migration is limited by inefficient (and exogenously given) labor allocation rules such as farm workers being paid average rather than marginal product or artificially limited absorption in the urban sector (e.g., formal sector minimum wages). The literature focuses on the effect on migration from the rural to urban sector of policies such as rural-urban terms of trade, migration

restrictions, wage subsidies, and the like.

The final and most complex versions of dual sector models are in Kelley and Williamson (1984) and Becker, Mills, and Williamson (1992), which are fully dynamic CGE models. They have savings behavior and capital accumulation, population growth, and multiple economic sectors in the urban and rural regions. Labor markets within sector and across regions are allowed to clear. The models analyze the effects of a wider array of policy instruments, including sector specific trade or capital market policies for housing, industry, services and the like. However the starting point is again an exogenously given initial urban-rural productivity gap, sustained initially by migration costs and exogenous skill acquisition. On-going urbanization is the result of exogenous forces -- technological change favoring the urban sector or changes in the terms of trade favoring the urban sector.

As models of urbanization, these dual economy ones are a critical step but they suffer obvious defects. First how the dual starting point arises is never modeled. Second, and related to the first, there are no forces for agglomeration that would naturally foster industrial concentration in the urban sector. Finally although the models have two sectors there is really little spatial or regional aspect to the problem. There is a new generation of two-sector models, the core-periphery models, which attempt to address some of these defects. The core-periphery models ask under what conditions in a two-region country, industrialization, or "urbanization" is spread over both regions versus concentrated in just one region.

Compared to the dual economy models, Krugman's (1991a) paper explicitly has scale economies that foster endogenous regional concentration. Second, while there are two regions, no starting point is imposed, where one region is assumed to start off ahead of the other. Industrialization may occur in both regions or in only one region. One region can become "backward" (under certain assumptions), or, if not backward (lower real incomes) at least relatively depopulated (Puga, 1999). But these are outcomes solved for in the model. Third the models have some notion of space represented as transport costs of goods between regions.

The models are focused on a key developmental issue -- the initial development of a core (say, coastal) region and a periphery (say, hinterland) region, as technology improves (transport costs fall) from a situation starting with two identical regions. As such they do relate to the earlier discussion in Section 1.3 of urban concentration in a primate city versus the rest of the urban sector. Some work (Puga 1999, Fujita, Krugman, and Venables 1999, Chapter 7, Helpman 1998, and Tabuchi 1998) also analyzes how under certain conditions, with further technological improvements, there can be reversal. Some industrial resources leave the core; and the periphery also industrializes/urbanizes. However core-periphery models have limited implications for urbanization per se, since in many versions including Krugman's (1991a) initial paper, the agricultural population is fixed.

Unfortunately, to date core-periphery models have been almost exclusively uni-dimensional in focus, asking what happens to core-periphery development as transport costs between regions decline. They are not focused on other forms of technological advance, let alone endogenous technological development. With a few exceptions, Fujita and Thisse (2002) and Baldwin (2001), the models are static. But even in these exceptions, there is still the focus on exogenous changes in transport technology. Compared to the older dual economy literature, generally core-periphery models have no policy considerations of interest to development economists, such as the impact of wage subsidies, rural-urban terms of trade, capital market imperfections. An exception is that some papers have examined the impact on core-periphery structures of reducing barriers to international trade, such as tariff reduction, and papers are starting to explore issues of capital market imperfections. The core-periphery model is an important innovation in bringing back the role of transport costs, largely ignored in urban systems work, to the forefront. Excellent summaries of the key elements include Neary (2001), Fujita and Thisse (2000) and Ottaviano and Thisse (2004), with the latter two developing many extensions. Fujita, Krugman and Venables (1999) stands as a basic reference on detailed modeling.

The dual economy and core-periphery models are regional models, with limited urban implications. Urban models are focused on the city formation process, where the urban sector is composed of numerous cities, endogenous in number and size. Efficient urbanization and growth require timely formation of cities. As policy issues the extent of market completeness in the national markets in which cities form, the role of city governments and developers, the role of inter-city competition, and the role of debt finance and taxation are critical. In the next section we analyze an urbanization process in which there are cities. Then we turn to a discussion of some key policy issues.

3.2 Urbanization With Cities

Here I present a simple two sector model of urbanization with cities, adapting the model in part 2 following Henderson and Wang (2005, 2004). The urban sector is exactly like the X_1 type of city sector earlier with production technology given in (6). The other sector is food produced in the agriculture sector, which we make now the numeraire (since there may initially be no urban sector). As a result, for type 1 cities in the urban sector, equation (7) for wages, equations (9) and (10) for commuting costs and rents, and equation (15) for income are all redefined to be multiplied by the price of X_1, p. The city size equation is the same, invariant to relative prices. Critical here is

$$I_1 = W_1 \ (1 - 2\delta_1) = pQ_1 h_1^{\varepsilon_1} \qquad (15^1)$$

for Q_1 a parameter cluster.

Agriculture. Rural output per worker is $D_a h_a^{\psi_a} h_a^{\theta_a}$, or $D_a h_a^{\varepsilon_a}$, so rural wages and real income are

$$W_a = D_a h_a^{\varepsilon_a} \qquad \varepsilon_a \equiv \psi_a + \theta_a \tag{20}$$

As such the rural sector is very simple: no commuting costs, no agglomeration economies and no diminishing returns to land. As in the urban sector, productivity is affected by individual human capital accumulation, $h_a^{\theta_a}$, and by sector knowledge spillovers, $h_a^{\psi_a}$.

Preferences and Urbanization. To have sectoral transformation we need to move away from the world of unitary price and income elasticities in Section 2, so growth between sectors is not parallel. Here we assume preferences have the form

$$V = (x + a^{\gamma})^{\alpha}, \qquad \gamma, a < 1 \tag{21}$$

where *a* is consumption of agricultural products. In (21) agricultural demand is income inelastic, with a demand function

$$a = \gamma^{\frac{1}{1-\gamma}} \rho^{\frac{1}{1-\gamma}}. \tag{22}$$

3.2.1 Human Capital Market, Migration, Savings.

The urbanization process as a "transitory" phenomenon is not a steady state process. To simplify, following much of the literature, we introduce an explicit market for human capital, as though human capital investments were not embodied. And we assume an exogenous savings rate, s. For the former, now each person in the economy has a human capital level, h, which can be used in production or can be loaned out. I now flesh out the equations of the model, that in Section 2 were skimmed over. The capital market equalizes capital returns across sectors so $r = p\ \theta_1\ D_1\ h_1^{\theta_1+\psi_1-1}\ n_1^{\delta_1} = \theta_a\ D_a\ h_a^{\theta_a+\psi_a-1}$. Substituting in for n_1 from (13)

$$p = Q_2\ h_a^{\varepsilon_a} h_1^{1-\varepsilon_1} \tag{23}$$

recalling $\varepsilon_a \equiv \psi_a + \theta_a$ and $\varepsilon_1 \equiv (\theta_1 + \psi_1)/(1 - 2\delta_1)$. Q_2 is a parameter cluster.

Free migration requires urban incomes defined to be gross human capital returns to equal the same for agriculture, or $I_1 + r\ (h - h_1) = W_a + r\ (h - h_a)$. Utilizing (15)

$$W_1 - W_a = pbn^{1/2} + r\ (h_1 - h_a) \tag{24}$$

With free migration equalizing real incomes across sectors, urban wages exceed rural wages by (commuting) cost-of-living differences (the first term on the RHS of (24)), and by a factor compensating if human capital requirements in the urban sector exceed those in the rural, as I assume.

If we substitute in (24) for $W_1,\ W_a,\ p$ and r and rearrange, we get

$$h_a = h_1\ \frac{\theta_a}{\theta_1}\ \left(\frac{1-\theta_1-2\delta}{1-\theta_a}\right) \tag{25}$$

A sufficient condition for $h_1 > h > h_a$, or the urban sector to be human capital intensive, is that $\theta_1 > \theta_a$, as assumed.

To close the model requires three relationships. First is national full employment of capital and labor so

$$n_a h_a + n_1 m_1 h_1 = hN \tag{26a}$$

$$n_a + m_1 n_1 = N \tag{26b}$$

where m_1,as before, is the number of type 1 cities and N is the national population. The third equation equates the demand for food equal to its supply. But that requires a digression on how human capital is produced and nature of savings. Since we want to be able to start with a purely rural economy, we don't want to have it produced just from X_1 as in Section 2.

We assume human capital production in each sector is made from goods from that sector (where an equal expenditure in any sector results in the same human capital), which is almost like assuming, for a fixed savings rate, a fixed fraction of working time in any sector is needed to produce a unit of human capital. Second, we assume savings at the rate s are from wage income net of rental costs, or from $I_1 - rh_1,$ and $W_a - rh_a$, which magnitudes are equalized by migration. Thus in the food market total production $n_a D_a h_a^{\varepsilon_a}$ equals food consumption demand $N\ (\rho\gamma)^{\frac{1}{1-\gamma}}$ (see 22) plus agricultural savings,

or $n_a D_a h_a^{\varepsilon_a} = N\ (p\gamma)^{\frac{1}{1-\gamma}} + sn_a\ (W_a - rh_a)$. Substituting in for r, for p from (22), W_a from (20) and for h_a from (25) we get

$$n_a / N = Q_3 h_1^{\frac{\gamma\varepsilon_a - \varepsilon_1}{1-\gamma}} \tag{27}$$

with Q_3 a parameter cluster. We assume $\gamma\varepsilon_a - \varepsilon_1 < 0$, so the social returns to human capital in the urban sector exceed those in the rural sector discounted by γ. With economic growth in human capital, the rural sector diminishes. Note in (27) for there to be an urban sector, h_1 must be large enough so $n_a / N < 1$, as we explain below. Of course h_1 is linked to h through (26a) where with substitutions

$$h_1 \left(1 - Q_4\ h_1^{\frac{\gamma\varepsilon_a - \varepsilon_1}{1-\gamma}}\right) = h \tag{28}$$

where given $\gamma\varepsilon_a - \varepsilon_1 < 0,\ dh_1 / dh > 0$, once there is an urban sector. Q_4 is a parameter cluster.

3.2.2 Urban Growth and Transformation

Once an urban sector exists, city growth is as in section 2: $\dot{n}_1 / n_1 = 2\varepsilon_1 \dot{h}_1 / h_1$, so cities grow with human capital accumulation. The growth in number of cities now depends on the rate of urbanization, as well. Combining (26a) and (26b), with differentiation $\dot{m}_1 / m_1 = (N / m_1 n_1)\ g - \dot{n}_1 / n_1 - (n_a / m_1 n_1)\ \dot{n}_a / n_a$. If we differentiate (27) for $\dot{n}_a / n_a$ and combine this becomes

$$\dot{m}_1 / m_1 = g - \dot{n}_1 / n_1 + \frac{n_a}{m_1 n_1}\ \frac{\gamma\varepsilon_a - \varepsilon_1}{1-\gamma}\ \dot{h}_1 / h_1 \tag{29}$$

As before the rate of growth of numbers of cities is increased by national population growth, g, and reduced by growth in individual city sizes. Now it is also enhanced by economic growth which increases relative demand for urban products and draws labor out of agriculture, as captured by the last term in (29).

3.2.3 Economic Growth

Given the savings rule, total human capital increases by $\dot{H} = s\left[m_1 n_1\ (I_1 - rh_1) + n_a\ (W_a - rh_a)\right]$ each instant so per person change in capital is $\dot{h} / h = \dot{H} / H - g$. Given $I_1 - rh_1 = W_a - rh_a$, with substitutions we have

$$\dot{h}/h = sQ_5 h_1^{\varepsilon_a - 1} \left(1\text{-}Q_6 h_1^{\frac{\gamma\varepsilon_a - \varepsilon_1}{1-\gamma}} \right)^{-1} - g \qquad (30)$$

where $Q_6 < Q_4$, for parameter clusters. In terms of growth, if $\varepsilon_a < 1$ and urbanization occurs, we have steady state levels given $\dot{h}/h$ declines with increases in h_1, and hence *h*. If $\varepsilon_a = 1$, we approach steady state growth once h_1 gets large so $n_a N \to 0$ and the expression in parentheses in (30) approaches 1. However in either case at low levels of development in (27) n_a / N is bounded at 1 where equation (27) defines a critical h_1 and hence h in (28), say h_c, below which $n_a / N = 1$. To have steady state levels with urbanization given we start at $n_a / N = 1$ with $\dot{h}/h = \dot{h}_a / h_a = -g + (1-\theta_a)\, h_a^{\varepsilon_a - 1}$, requires $h_c < (g/(1\text{-}\theta_a))^{\frac{1}{\varepsilon_a - 1}}$, so we pass the critical h at which urbanization starts before hitting the potential steady state value of *h* without urbanization. Otherwise the economy can be stuck with no urbanization. Details of this and issues of multiple equilibria are discussed in Henderson and Wang (2005).

3.3 Extensions and Policy Issues

There are three general sets of policy issues. First concerns whether in the context of the models in section 2 and 3.2, the national composition of cities of different types is efficient. We have already discussed this issue: in many contexts asking whether the national composition of cities is efficient is like asking if national output composition is efficient. If there are national policy biases such as trade policies favoring steel products over textile products, with urban specialization, if steel is produced in bigger types of cities than textiles, the numbers of larger cities relative to smaller ones and hence urban concentration will increase. The second set of policy issues concerns whether, in general, city sizes are likely to be efficient and we discuss this in Section 3-3-1.

The second general set of issues deals with factors we have ignored. In particular the modeling in sections 2 and 3.2 assumes a nice smooth process where (i) all factors of production are perfectly mobile and malleable, (ii) city borrowing and debt accumulation have no role, (iii) "lumpiness" problems that arise in city formation when economies are small are ignored: while m must be an integer in reality, in the analysis it is treated as any positive number where the number of cities grows at a rate $\dot{m}/m$, rather than by 0, 1 or 2. A model that incorporates these features is outlined in section 3.3.2, which brings to the forefront a variety of policy issues.

3.3.1 City Sizes

A perpetual debate in particular developing countries is whether certain mega-cities are oversized, squandering national resources that must be allocated to commuting, congestion, and transport

in those cities and resulting in low quality of life in the polluted, unsanitary and crowded slums of such cities. In other countries especially former planned economies the debate goes the other way: are cities too small? The growth connection is straightforward. Either squandered resources in over-sized cities or too small cities with unexploited scale economies mean lower income levels, potentially lower savings, lower capital accumulation and thus lower growth rates. While calculations are tedious, in the steady state growth in section 2.2.2 where $\gamma^h = \dot{h}/h = (A-\rho)/\sigma$, A depends on urban parameters (for example, increasing with human capital returns) and will be lowered if city sizes are inefficient.

Using a simple, partial equilibrium diagram, it is possible to illustrate both issues: the mega-city "problem" and the planned economy problem. The diagrams point to first order effects. For the mega-city problem, suppose there are a variety of type 1 cities in an economy with free mobility of labor and institutions supporting efficient city formation. In Figure 5a, the representative city has a size n_1^*, where real income as function of city size peaks at I_1^*, tangent to the perfectly elastic national supply curve of labor to the city. Suppose one particular type 1 city is favored relative to the rest, where the various types of favoritism are discussed in Section 1.3.2. For example it may have special public services compared to other cities financed out of national taxes. Those favors raise the utility, or real income that residents in the favored city potentially receive, shifting up the inverted-U real income curve. That upward shift draws migrants into the city expanding its size to n_{mega}. But at n_{mega}, the net income generated by the city, ignoring its nationally financed favors, is only net I_m. The gap, $I_1^* - I_m$, times the population represents "squandered resources". Of course, such squandering would in general equilibrium affect prices, lowering the height of the population supply curve and the inverted-U's.

A second issue in city formation concerns poor institutions in national land markets and in local governance which limit the number of cities that can form. Suppose that, in villages which might become cities, local governments by institutional restrictions can't expand infrastructure (see next section), can't rezone and build on urban fringe land, and can't offer subsidies to incoming firms. And suppose developers can't assemble large tracts of land for development because property rights are ill-defined. These villages can't grow into cities; nor can entirely new cities form. If the number of cities is bindingly limited, so there are too few cities, all existing cities under free migration are too big. In Figure 5a, suppose we reconsider the figure ignoring the representative city curve and assume all cities have inverted-U's like the favored city. Then, in this reinterpretation of the figure, I_F is the potentially attainable real income in all cities (ignoring general equilibrium effects) if cities could freely form. Given restricted numbers, rather than operating at I_F (with size n_1^*), cities are overcrowded; and in equilibrium

they operate at I_1^* (with size n_{mega}), with the same national supply curve of labor as labeled in the figure. The restrictions result in losses related to the gap $I_F - I_1^*$.

The planned economy problem is entirely different. Former "planned" economies like China have formal migration restrictions limiting the visas given for rural people to move to cities and limiting migrants access to jobs, housing, medical care and schooling in destination cities to reduce the incentive to migrate. Some former planned economies (as well as China) limited migration through housing provision and land development. If the state provides and allocates all housing assignments, migrants can't move unless housing is provided in the destination. As we saw in Table 2, countries like China and Russia have very low urban concentration compared to other large countries. Figure 5b captures the essence of the problem. While the representative city has an inverted-U where real income is maximized at n_1^*, migration restrictions for cities a and b restrict sizes to n_a and n_b and real incomes to I_a and I_b. Au and Henderson (2002) estimate these inverted-U's for different types of cities in China in 1997 and find that 30% of cities are significantly undersized – below the lower 95% confidence interval on their equivalent to n_1^*. The productivity losses from being undersized are enormous: 30-50% or more loss in GDP per capita for many cities.

3.3.2 Sequential City Formation and Governance

In a working paper, Henderson and Venables (2004) take a new approach to city formation. They assume a context where (1) there is a steady-flow of migrants from rural to urban areas, and (2) urban residence requires a fixed investment in non-malleable, immobile capital (housing, sewers, water mains, etc.). Cities form sequentially without population swings, so migrants all flow first into city 1 until its equilibrium size is reached (abstracting from any on-going technological change), and then all future migrants all go to a second city until its equilibrium size is reached, and so on. This is a very different process than when all resources are mobile: in the usual models in a small economy, when the second city forms, it takes half the population of the first at that instant, and when the third forms it takes one third of the then population of the first two. Cities grow way past $\overset{*}{n}_1$, shrink back to $\overset{*}{n}_1$, and then grow again, shrink, and on so. With fixed capital, such population swings would mean periods of abandoned housing. With sufficiently high required capital investments all population swings are eliminated in equilibrium. Each new city starts off tiny with no accumulated scale effects and low productivity. It grows steadily absorbing all new rural-urban migrants until its growth interval is complete and it reaches steady state size; then a new city starts off growing from a tiny size.

With sequential city formation without population swings, given discounting of the future, efficient city size requires cities to grow past the equivalent of $\overset{*}{n}_1$ to their steady state size, n_{opt}, at which

point real income per worker is declining. Intuitively, growing past n_1^*, with declining but still high real income, postpones the formation of a new city with tiny population, no scale effects and very low incomes. The paper then looks at equilibrium city formation in two contexts.

First is a situation with no "large agents" in national land markets – no developers and no city governments. In a model with perfect mobility of resources as discussed in Section 2, city formation with atomistic agents is a disaster due to coordination failure. A new city can only form when old cities are so big that the income levels they offer have fallen to the point where they equal what a person can earn in a city of size one. Having immobile capital presents a commitment device (Helsley and Strange, 1996), so individual, sequentially rational builders switch from building in an old city to building in a new one at a "reasonable time". Real incomes are still equalized across cities through migration. Given big old cities have high nominal incomes and the tiny new one low nominal income, housing rents adjust in old cities to equalize real incomes. Housing rents in old cities change over the growth cycle of a new city, starting very high and then declining (see also Glaeser and Gyourko, 2003). In this context, equilibrium sizes may even be smaller than optimal ones. The deviation from optimum has not to do with coordination failure which is solved despite the absence of "large" agents, but with the present value of externalities from a migrant in an old versus a new city.

With developers or full empowered local governments, externalities are appropriately internalized and city sizes are n_{opt}. However, apart from financing the housing and infrastructure capital, to induce new migrants to move to a new city with its low real income and scale economies in a timely fashion, large agents must subsidize in-migration of worker-firms. To do this they must borrow and, in fact, public debt accumulates over the entire growth interval of a city and only starts to be paid off once it reaches steady state size. Debt ceilings, or limits for cities which are common in many countries curtail subsidies to in-migrants and postpone new city formation. Debt limited cities are too big. The paper also explores the effects of limits on local tax property tax powers.

4. Some Issues For a Research Agenda

A handbook paper is a place to offer research suggestions, as well as summarize the state of knowledge. While various avenues of needed research are noted throughout, here I summarize three key suggestions. In all the spatial and urban work, transport costs are either absent or treated as a technology parameter that may exogenously change. In an actual development and growth context, transport costs reflect public infrastructure investment decisions, subject to political influence. Core-periphery models need to endogenize transport costs and urban models consider them, so spatial structures across regions

and cities are an outcome of investment decisions. A similar comment involves mobility costs of workers, which are related to both transport and communication infrastructure investments.

A second key research issue involves spatial inequality as it evolves with growth in a context where workers have different ability endowments and choose different human capital levels. In most spatial and urban models, workers are identical, except for their degree of mobility. But with urbanization, it may be that it is higher ability rural folks who urbanize and acquire modern skills, increasing real income gaps between high and low ability people. We have no models that directly address these issues and provide a comprehensive framework to evaluate spatial inequality, or cross-space income differences.

Finally we don't really have models that address the evolution of city production patterns with on-going technological change. While we have looked at parallel growth and urbanization, we know city functions also change over time. In less developed countries, bigger cities may be focused on manufacturing, but somehow with growth and technological change, big cities tend to specialize more in service functions, purchased by manufacturers and retailers in smaller cities. While we have models of functional specialization , we haven't modeled this evolution in city roles over the development process.

References

Abdel-Rahman, H. (1996), "When do Cities Specialize in Production," Regional Science and Urban Economics, 26 1-22.

Abdel-Rahman, H. (2000), "City Systems: General Equilibrium Approaches," in J-M Huriot and J-F Thisse (eds.), Economics of Cities: Theoretical Perspectives, Cambridge University Press, 109-37.

Abdel-Rahman, H. and A. Anas (2004), "Theories of Systems of Cities," in J.V. Henderson and J-F. Thisse (eds.), Handbook of Urban and Regional Economics, Vol. 4, Cities and Geography, North Holland.

Abdel-Rahman, H. and M. Fujita (1990), "Product Variety, Marshallian Externalities, and City Sizes," Journal of Regional Science, 30, 165-85.

Abdel-Rahman, H. and P. Wang (1997), "Social Welfare and Income Inequality in a System of Cities," Journal of Urban Economics, 41, 462-83.

Ades, A.F. and E.L. Glaeser (1995), "Trade and Circuses: Explaining Urban Giants," Quarterly Journal of Economics, 110, 195-227.

Anas, A. and K. Xiong, (1999), "The Formation and Growth of Specialized Cities," State University of New York at Buffalo mimeo.

Arthur (1990), "'Silicon Valley Locational Clusters: When Do Increasing Returns to Scale Imply Monopoly," Mathematical Social Sciences, 19, 235-51.

Au, C.C. and J.V. Henderson (2002), "How Migration Restrictions Limit Agglomeration and Productivity in China," NBER Working Paper No. 8707.

Baldwin, R.E. (2001), "Core-Periphery Model with Forward-Looking Expectations," Regional Science and Urban Economics, 31, 21-49.

Barro, R. and X. Sala-i-Martin (1991), "Convergence Across States and Regions," Brookings Papers on Economic Activity, 1,107-82.

Barro R. and X. Sala-i-Martin (1992), "Regional Growth and Migration: A Japan-USA Comparison," Journal of Japanese and International Economics, 6, 312-46.

Becker, G., E. Mills, J.G. Williamson (1992), Indian Urbanization and Economic Growth Since 1960, Johns Hopkins Press.

Beeson, P.E., D.N. DeJong, and W. Troeskan (2001), "Population Growth in US Counties, 1840-1990," Regional Science and Urban Economics, 31, 669-700.

Benabou, R. (1993), "Workings of a City: Location, Education, and Production," Quarterly Journal of Economics, 108, 619-52.

Bergsman, J., P. Greenston, and R. Healy (1972), "The Agglomeration Process in Urban

Growth," Urban Studies, 9, 263-88

Black, D. (2000), "Local Knowledge Spillovers and Inequality," University of California Irvine mimeo.

Black, D. and J.V. Henderson (1999a), "A Theory of Urban Growth, " Journal of Political Economy, 107, 252-84.

Black, D. and J.V. Henderson (1999b), "Spatial Evolution of Population and Industry in the USA," Papers and Proceedings of the American Economic Association, May.

Black, D. and J.V. Henderson (2003), "Urban Evolution in the USA," Journal of Economic Geography, 3,343-373.

Clark, J.S. and J.C. Stabler (1991), "Gibrat's Law and the Growth of Canadian Cities," Urban Studies, 28, 635-39.

Cordoba, J-C (2004), "On the Size Distribution of Cities," Rice University mimeo.

Davis, James (2000), "Headquarter Service and Factory Urban Specialization With Transport Costs," Brown University mimeo.

Davis, J. and J.V. Henderson (2003), "Evidence on the Political Economy of the Urbanization Process," Journal of Urban Economics, 53, 98-125.

Dixit A. and J. Stiglitz (1977), "Monopolistic Competition and Optimum Product Diversity," American Economic Review, 67, 297-308.

Dobkins, L.H. and Y.M. Ioannides (2001), "Spatial Interactions Among U.S. Cities: 1900-1990," Regional Science and Urban Economics, 31, 701-32.

Duranton, G. (2004), "City Size Distribution as a Consequence of the Growth Process," LSE mimeo.

Duranton, G. and H. Overman (2004), "Testing for localization using micro-geographic data," Revised working paper, April 2004.

Duranton, G. and D. Puga (2000), "Nursery Cities: Urban Diversity Process Innovation, and the Life Cycle of Products," American Economic Review, 91, 1454-77.

Duranton, G. and D. Puga (2001), "From Sectoral to Functional Urban Specialization", CEPR, LSE Discussion Paper 2971.

Duranton, G. and D. Puga (2004), "Microfoundations of Urban Agglomeration Economies" in Handbook of Urban and Regional Economics, Volume 4, J.V. Henderson and J-F Thisse (eds.), North Holland.

Durlauf, S.N. (1996), "A Theory of Persistent Income Inequality," Journal of Economic Growth, 1, 75-93.

Eaton, J. and Z. Eckstein (1997), "Cities and Growth: Evidence from France and Japan,"

Regional Science and Urban Economics, 27, 443-74.

Ellison, G. and E. Glaeser (1999a), "The Geographic Concentration of US Manufacturing: A Dartboard Approach," Journal of Political Economy, 105, 889-927.

Ellison, G. and E. Glaeser (1999b), "The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration," American Economic Association Papers and Proceedings, 89, 311-16.

Fay, M. and C. Opal (1999), "Urbanization Without Growth: Understanding an African Phenomenon," World Bank mimeo.

Flatters, F., J.V. Henderson and P. Mieszkowski (1974), "Public Goods, Efficiency, and Regional Fiscal Equalization," Journal of Public Economics, 3, 99-112.

Fujita, M. and H. Ogawa (1982), "Multiple Equilibria and Structural Transition of Non-Monocentric Configurations," Regional Science and Urban Economics, 12, 161-96.

Fujita, J., P. Krugman and A.J. Venables (1999), The Spatial Economy: Cities, Regions, and International Trade, MIT Press.

Fujita, M. and T. Ishii (1994), "Global Location Behavior and Organization Dynamics of Japanese Electronic Firms and Their Impact on Regional Economies," Paper presented for Prince Bertil Symposium on the Dynanmic Firm, Stockholm.

Fujita, M. and J-F Thisse (2000), "The Formation of Economic Agglomerations," in J-M Huriot and J-F Thisse (eds.) Economies of Cities, NY, Cambridge University Press.

Fujita, M. and J-F Thisse (2002), Economics of Agglomeration, Cambridge University Press.

Gabaix, X. (1999a), "Zipf's Law and the Growth of Cities," American Economic Association and Proceedings, 89, 129-32.

Gabaix, X.(1999b), "Zipf's Law for Cities: an Explanation," Quarterly Journal of Economics, 114, 739-67.

Gallup, J.L., J.D. Sacks and A. Mellinger (1999), "Geography and Economic Development," International Regional Science Review, 22, 179-232.

Galor, O. and J. Zeira (1993), "Income Distribution and Macro Economics," Review of Economic Studies, 60, 35-52.

Glaeser, E., J. Scheinkman and A. Schelifer (1995), "Economic Growth in a Cross-Section of Cities," Journal of Monetary Economics, 36, 117-34.

Glaeser, E. and J. Gyourko (2003), “Urban Decline and Durable Housing,” Harvard University mimeo.

Grossman, G. and E. Helpman (1991), "Quality Ladders in the Theory of Growth," Review of Economic Studies, 58, 43-61.

Harris, J. and M. Todaro (1970), "Migration, Unemployment and Development: A Two Sector Analysis," American Economic Review, 40, 126-42.

Head, K and T. Mayer (2004), "The Empirics of Agglomeration and Trade[*]," in J.V. Henderson and J-F Thisse (eds.), Handbook of Urban and Regional Economics, Vol. 4, Cities and Geography, North Holland.

E. Helpman (1998), "The Size of Regions," in D. Pines, E. Sadka and I. Zilcha (eds.), Topics in Public Economics: Theoretical and Applied Analysis, Cambridge University Press, 33-54.

Helsley, R. and W. Strange (1990), "Matching and Agglomeration Economies in a System of Cities," Regional Science and Urban Economics, 20, 189-212.

Helsley, R. and W. Strange (1994), "City Formation with Commitment," Regional Science and Urban Economics, 24, 373-390.

Henderson, J.V. (1974), "The Sizes and Types of Cities," American Economic Review, 61, 640-56.

Henderson, J.V. (1988), Urban Development: Theory, Fact and Illusion, Oxford University Press.

Henderson, J.V. (1997), "Medium Size Cities", Regional Science and Urban Economics, 27, 449-470.

Henderson, J.V. (2002), "Urban Primacy, External Costs, and Quality of Life," Resource Economics and Energy, 24, 95-106.

Henderson, J.V. (2003) "The Urbanization Process and Economic Growth: The So-What Question," Journal of Economic Growth, 8, 47-71.

Henderson, J.V. and Y. Ioannides (1981), "Aspects of Growth in a System of Cities," Journal of Urban Economics, 10, 117-39.

Henderson, J.V. and A. Kuncoro (1996), "Industrial Centralization in Indonesia," World Bank Economic Review 10, 513-40.

Henderson, J.V., A. Kuncoro and P. Nasution (1996), "Dynamic Development in Jabotabek," Indonesian Bulletin of Economic Studies, 32, 71-96.

Henderson, J.V. and R. Becker (2001), "Political Economy of City Sizes and Formation," Journal of Urban Economics, 48, 453-84.

Henderson, J.V., T. Lee and J.Y. Lee (2001), "Scale Externalities in Korea," Journal of Urban Economics, 49, 479-504.

Henderson, J.V. and A.J. Venables (2004), "The Dynamics of City Formation: Finance and Governance," LSE mimeo.

Henderson, J.V. and H.G. Wang (2004), "Urbanization and City Growth," Brown University mimeo.

Henderson, J.V. and H.G. Wang (2005), “Urbanization and City Growth,” Journal of Economic Geography, forthcoming

Hochman, O. (1977), "A Two Factor Three Sector Model of an Economy With Cities," mimeo.

Holmes, T. (1999), "Localization of Industry and Vertical Disintegration," Review of Economics and Statistics, 81, 314-25.

Ioannides, Y.M. and H.G. Overman (2003), “Zipf's Law for Cities: An Empirical Examination," Regional Science and Urban Economics, 33, 1, March, 127-137.

Junius, K. (1999), "Primacy and Economic Development: Bell Shaped or Parallel Growth of Cities," Journal of Economic Development, 24 (1), 1-22.

Kanemoto, Y. (1980), Theories of Urban Externalities, Amsterdam: North-Holland.

Kelly, A.C. and J.G. Williamson (1984), What Drives Third World City Growth? A Dynamic General Equilibrium Approach, Princeton University Press.

Kim, H.S. (1988), "Optimal and Equilibrium Land Use Pattern in a City: A Non-Parametric Approach," Ph.D. thesis, Brown University.

Kim, S. (1995), "Expansion of Markets and the Geographic Distribution of Economic Activities: The Trends in US Manufacturing Structure, 1860-1987," Quarterly Journal of Economics, 95, 881-908.

Kolko, J. (1999), "Can I Get Some Service Here? Information Technology Service Industries, and the Future of Cities," Harvard University mimeo.

Krugman, P. (1991a), "Increasing Returns and Economic Geography," Journal of Political Economy, 99, 483-99.

Krugman, P. (1991b), Geography and Trade, MIT Press, Cambridge.

Lee, K.S. (1988), "Infrastructure Constraints on Industrial Growth in Thailand," World Bank INURD Working Paper No. 88-2.

Lee, K.S. (1989), The Location of Jobs in a Developing Metropolis, Oxford University Press.

Lee, T.C. (1997), "Industry Decentralization and Regional Specialization in Korean Manufacturing," unpublished Brown University Ph.D. thesis.

Lewis, W.A. (1954), "Economic Development With Unlimited Supplies of Labor," Manchester School of Economic and Social Studies, 22, 139-91.

Lucas, R.E. (1988), "On the Mechanics of Economic Development," Journal of Monetary Economics, 12, 3-42.

Lucas, R.E. and E. Rossi-Hansberg (2002), "On the Internal Structure of Cities," Econometrica, 70:4, 1445-1476.

Marshall, A. (1890), Principles of Economics, London: MacMillan.

Mills, E. and C. Becker (1986), Studies in Indian Urban Development, Oxford University Press.

Mills, E. and B. Hamilton (1994), Urban Economics, Scott-Foresman.

Mohring, H. (1961), "Land Values and Measurement of Highway Benefits," Journal of Political Economy, 49, 236-49.

Neary, J.F. (2001), "Of Hype and Hyperbolas: Introducing the New Economic Geography," Journal of Economic Literature, 49, 536-61.

Ono, Y. (2000), "Outsourcing Business Service and the Scope of Local Markets," CES Discussion Paper CES 00-14.

Ottaviano, G, and J-F Thisse (2004), "Agglomeration and Economic Geography," in J.V. Henderson and J-F Thisse (eds) Handbook of Urban and Regional Economics, Vol. 4, Cities and Geography, North Holland.

Overman, H., S. Redding and A. J. Venables (2003), "The Economic Geography of Trade, Production and Income: A Survey of Empirics," Handbook of International Trade, J. Harrigan and K. Choi, (eds.), Blackwell.

Puga, D. (1999), "The Rise and Fall of Regional Inequalities," European Economic Review, 43, 303-34.

Quah, D. (1993), "Empirical Cross Section Dynamics and Economic Growth," European Economic Review, 37, 426-34.

Rannis, G. and J. Fei (1961), "A Theory of Economic Development," American Economic Review, 51, 533-65.

Rappaport, J. and D. Sacks (2003), "The US as a Coastal Nation," Journal of Economic Growth, 8, 5-46.

Rauch, J.E. (1993), "Does History Matter Only When It Matters a Little? The Case of City-Industry Location," Quarterly Journal of Economics, 108, 843-67.

Ray, D. (1998), Development Economics, Princeton: Princeton University Press.

Renaud, B. (1981), National Urbanization Policy in Developing Countries, Oxford University Press.

Rosen, K. and M. Resnick (1980), "The Size Distribution of Cities: An Examination of the Pareto Law and Primacy," Journal of Urban Economics, 81, 165-86.

Rosenthal, S. and W. Strange (2004), "Evidence on the Nature and Sources of Agglomeration Economies" in J.V. Henderson and J-F Thisse (eds.) Handbook of Urban and Regional Economics, Vol. 4, Cities and Geography, North Holland.

Rossi-Hansberg, E. (2004), "Optimal Urban Land Use and Zoning," Review of Economic Dynamics,7, 69-106.

Rossi-Hansberg, E. and E.M. Wright (2004), “Urban Structure and Growth”, Stanford University (May) mimeo

Simon, H. (1995), "On a class of skew distribution functions," Biometrika, 44, 425-40.

Stiglitz, J. (1977), "The Theory of Local Public Goods," in M.S. Feldstein and R.P. Inman (eds.), The Economics of Public Services, London: MacMillan, 273-334.

Tabuchi, T. (1998), "Urban Agglomeration and Dispersion: A Synthesis of Alonso and Krugman," Journal of Urban Economics, 44, 333-351.

Thomas, V. (1978), “The Measurement of Spatial Differences in Poverty: The Case of Peru,” World Bank Staff Working Paper No. 273.

Wheaton, W. and H. Shishido (1981), "Urban Concentration, Agglomeration Economies, and the Level of Economic Development," Economic Development and Cultural Change, 30, 17-30.

Williamson, J. (1965), "Regional Inequality and the Process of National Development," Economic Development and Cultural Change, June, 3-45.

World Bank (2000), Entering the 21st Century: World Development Report 1999/2000, Oxford University Press.

Xiong, K. (1998), "Intercity and Intracity Externalities in a System of Cities: Equilibrium, Transient Dynamics, and Welfare Analysis," unpublished Ph.D. thesis, State University of New York at Buffalo.

Table 1. World City Size Distribution, 2000

size range	**count**	**mean**	**Share[1]**
17,000,000 < =n2000	4	20,099,000	4.5
12,000,000 < =n2000 < 17,000,000	7	13,412,714	5.2
8,000,000 < =n2000 < 12,000,000	13	10,446,385	7.5
4,000,000 < =n2000 < 8,000,000	29	5,514,207	8.9
3,000,000 < =n2000 < 4,000,000	41	3,442,461	7.8
2,000,000 < =n2000 < 3,000,000	75	2,429,450	10.1
1,000,000 < =n2000 < 2,000,000	247	1,372,582	18.8
500,000 < =n2000 < 1,000,000	355	703,095	13.9
250,000 < =n2000 < 500,000	646	349,745	12.5
100,000 < =n2000 < 250,000	1,240	157,205	10.8
Overall	2,657	658,218	100.0

1) a ratio of total population in the group to total population of cities with > =100,000

Table 2. Spatial Inequality

	1960		2000		
	(1)	(2)	(3)	(4)	(5)
	Gini	Number of Cities	Gini	Number of Cities	Rank Size Coefficient "a"
World	.59	1197	.56	1673	n.a.
Developed	.65	523	.58	480	n.a.
Soviet bloc	.52	179	.45	202	n.a.
Less developed	.57	495	.56	991	n.a.
Brazil	.67	24	.65	65	-.87
China	.47	108	.43	223	-1.3
India	.56	95	.58	138	-1.1
Indonesia	.52	22	.61	30	-.90
Mexico	.61	28	.60	55	-1.04
Nigeria	.31	20	.60	38	-.98
France	.61	31	.59	27	.97
Germany	.6	44	.56	31	-.74
Japan	.60	100	.66	82	-1.06
Russia	.54	79	.46	91	-1.34
Spain	.53	27	.52	20	-.98
Ukraine	.44	25	.40	32	-1.31
UK	.68	39	.60	21	-.83
USA	.58	167	.54	197	-1.11

Table 3. Total Numbers of Cities and Sizes

	1960	**1970**	**1980**	**1990**	**2000**
number of cities	969	1,129	1,353	1,547	1,568
mean size	556,503	640,874	699,642	789,348	943,693
median size	252,539	275,749	304,414	355,660	423,282
minimum size	100,082	115,181	126,074	141,896	169,682

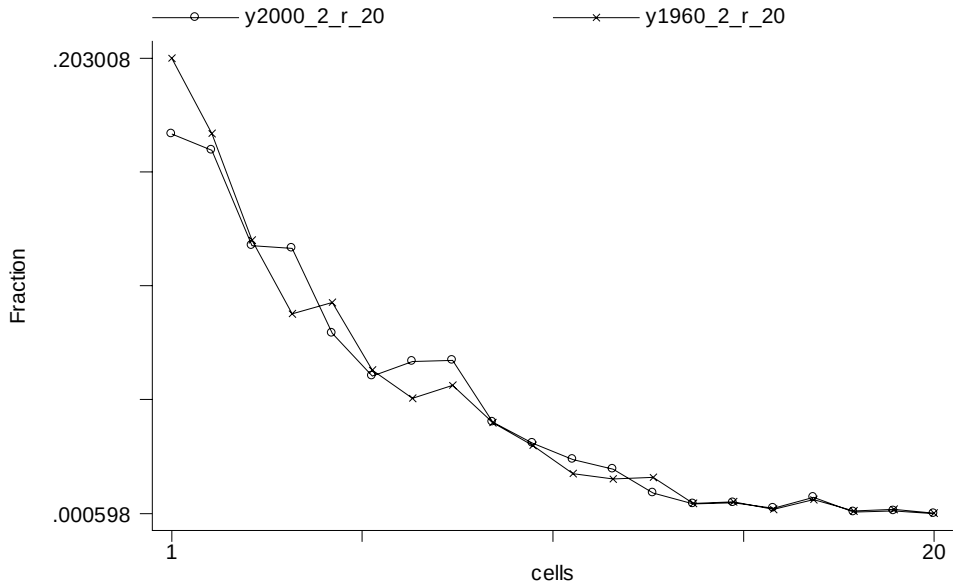


Figure 1a. Relative Size Distribution of Cities for all Countries

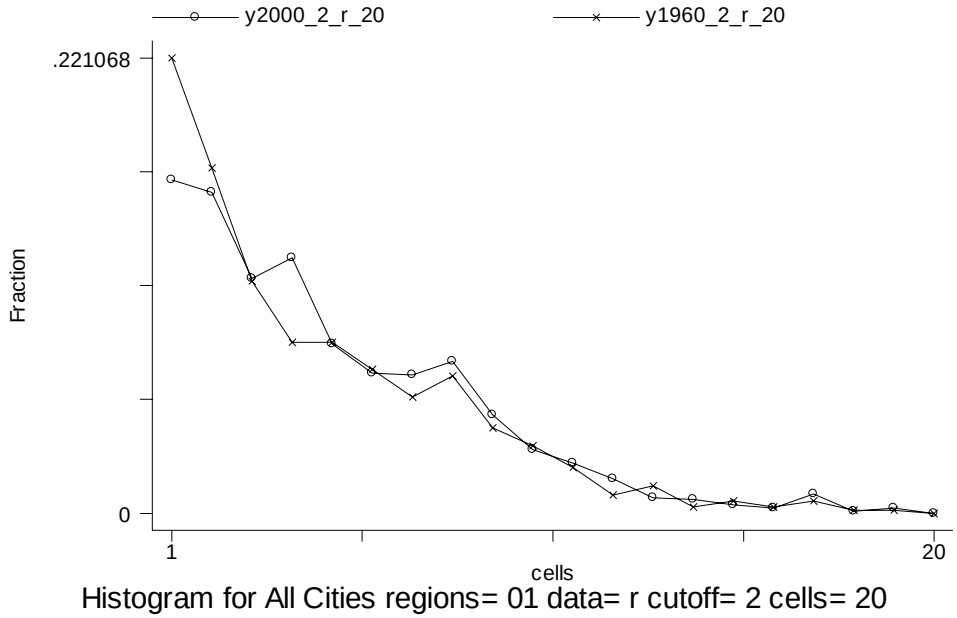


Figure 1b. Relative Size Distribution of Cities in Developing and Transition Economies

Figure 2: Share of Urban Population in Total Population.
(average over countries within groups)

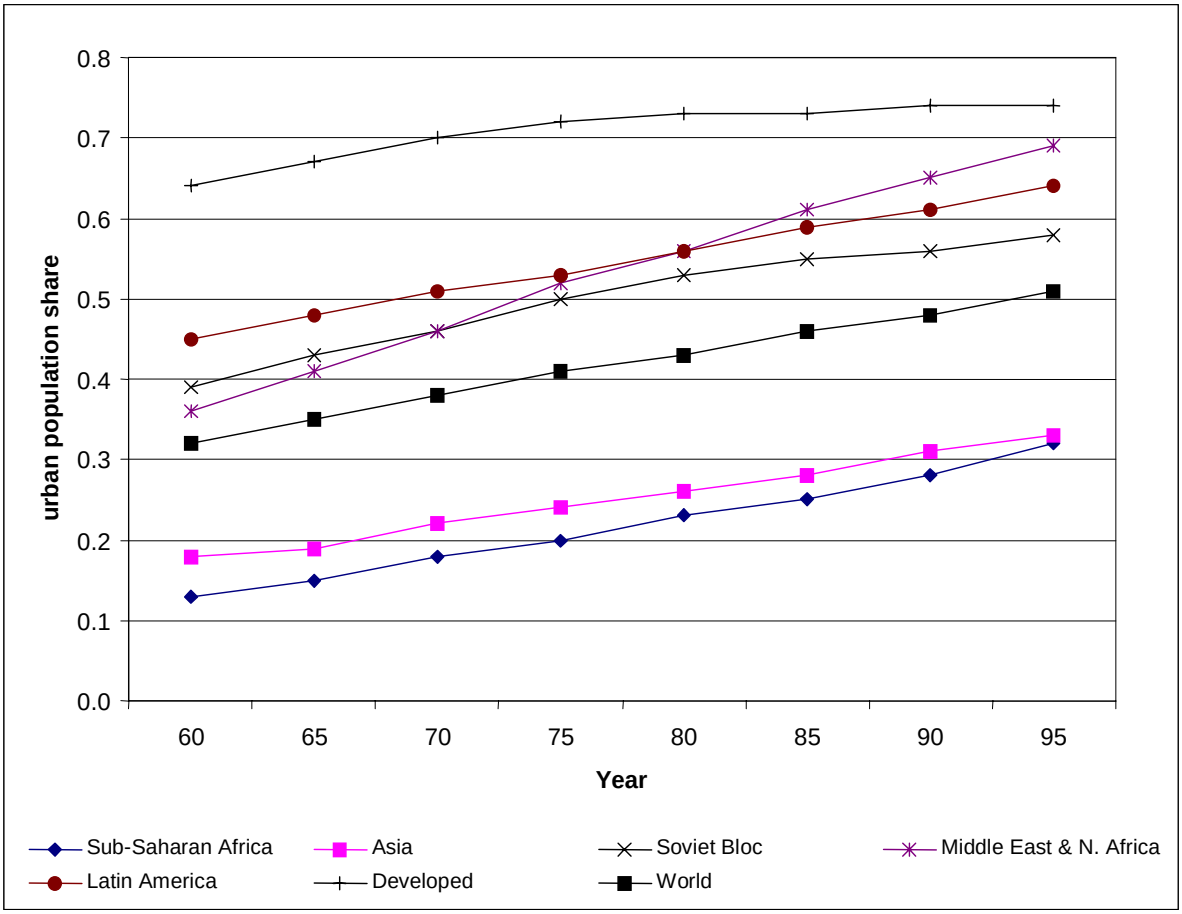


Figure 3: Partial Correlation Between Ln(urban population) and Ln(real GDP per capita), Controlling for Ln(national population), 1965-1995.

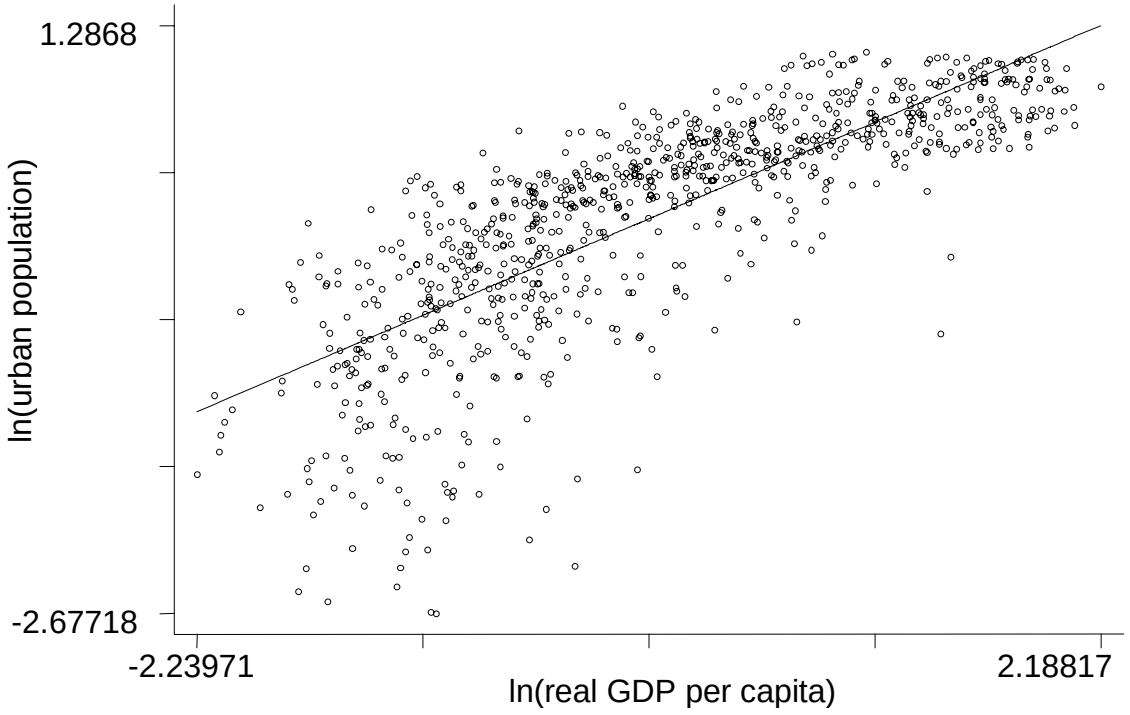
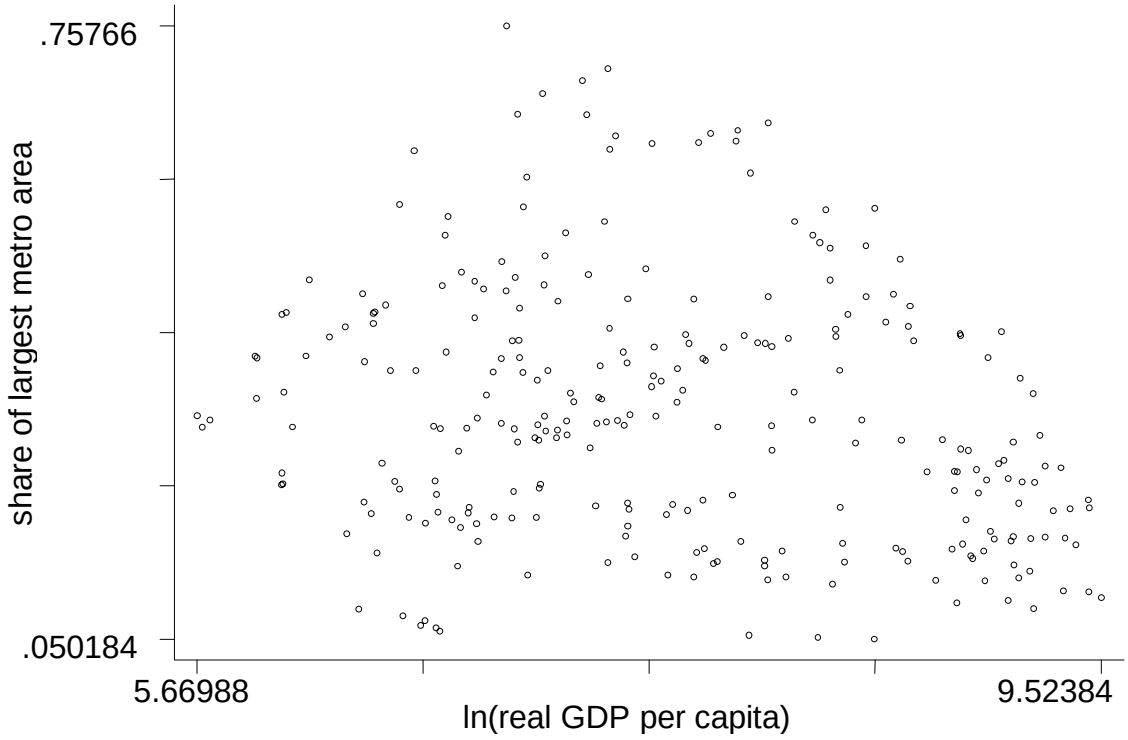


Figure 4: Primacy and Economic Development.

(a) Early period: 1965-75.



(b) Recent Period: 1985-95.

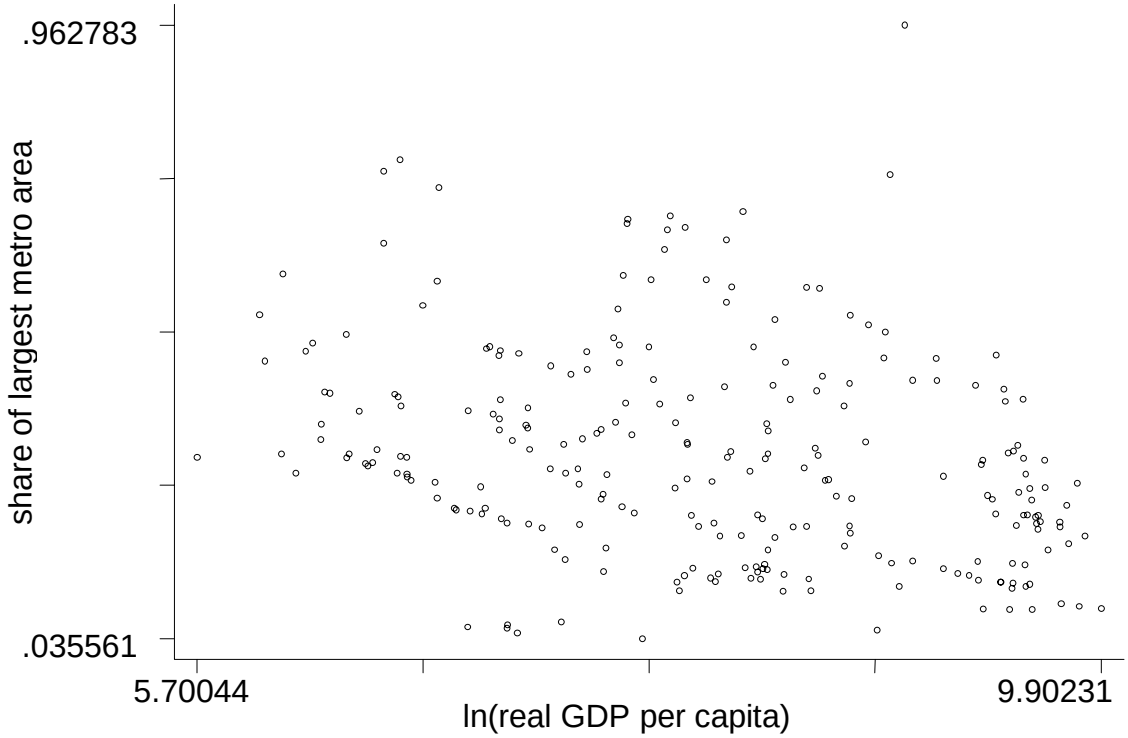
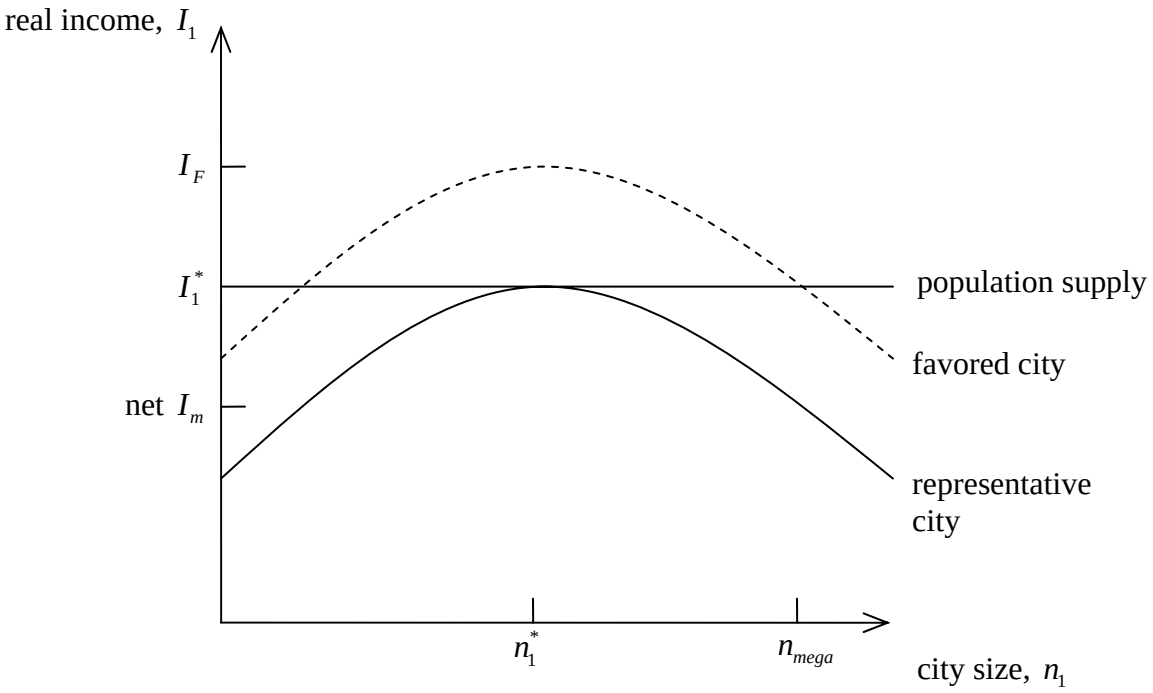
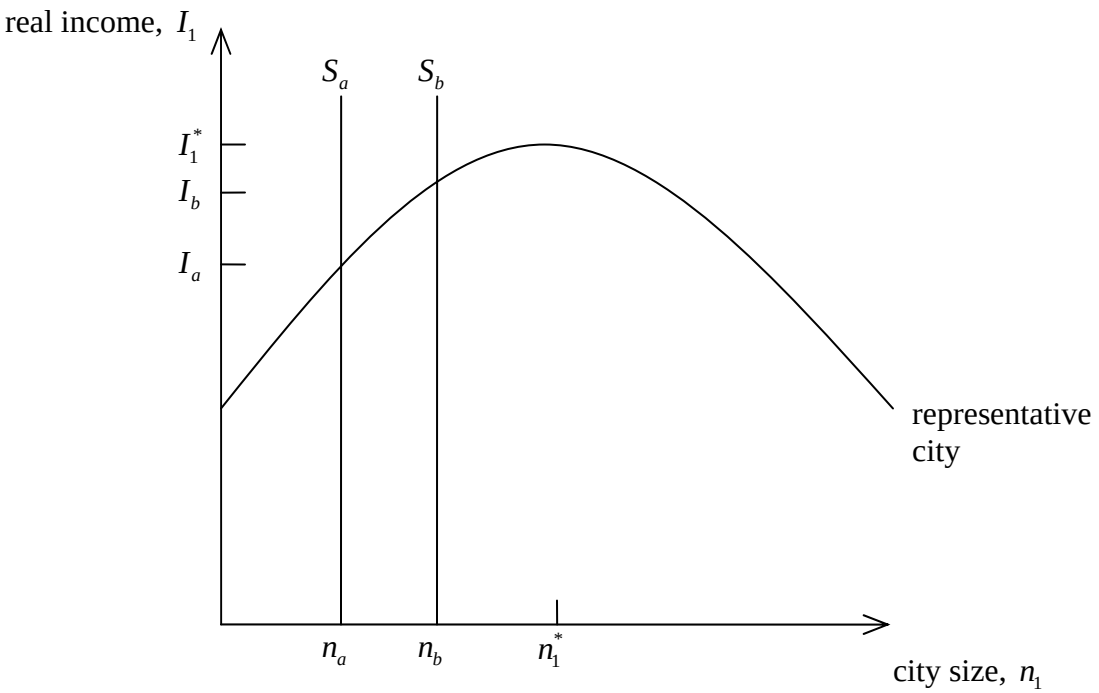


Figure 5. City Sizes



(a) Favored Cities



(b) Migration Restrictions

Inequality, Technology, and the Social Contract

Roland Bénabou[1]

First draft: May 2003

This version October 2004

[1]Princeton University, NBER, CEPR, IRP and BREAD. Forthcoming in the *Handbook of Economic Growth* (Philippe Aghion and Steven Durlauf, eds., North-Holland). I thank for useful remarks Philippe Aghion, Omer Moav and participants at the CEPR conference on Education and Inequality (Paris, May 2003). I am grateful to the MacArthur Foundation for financial support and to the Institute for Advanced Study for its hospitality over the academic year 2002-2003.

Abstract

The distribution of human capital and income lies at the center of a nexus of forces that shape a country's economic, institutional and technological structure. I develop here a unified model to analyze these interactions and their growth consequences. Five main issues are addressed. First, I identify the key factors that make both European-style "welfare state" and US-style "laissez-faire" social contracts sustainable.; I also compare the growth rates of these two politico-economic steady states, which are no Pareto-rankable. Second, I examine how technological evolutions affect the set of redistributive institutions that can be durably sustained, showing in particular how skill-biased technical change may cause the welfare state to unravel. Third, I model the endogenous determination of technology or organizational form that results from firms' tailoring the flexibility of their production processes to the distribution of workers' skills. The greater is human capital heterogeneity, the more flexible and wage-disequalizing is the equilibrium technology. Moreover, firms' choices tend to generate excessive flexibility, resulting in suboptimal growth or even self-sustaining technology-inequality traps. Fourth, I examine how institutions also shape the course of technology; thus, a world-wide shift in the technology frontier results in different evolutions of production processes and skill premia across countries with different social contracts. Finally, I ask what joint configurations of technology, inequality and redistributive policy are feasible in the long run, when all three are endogenous. I show in particular how the diffusion of technology leads to the "exporting" of inequality across borders; and how this, in turn, generates spillovers between social contracts that make it more difficult for nations to maintain distinct institutions and social structures.

Keywords: inequality, welfare state, technical change, skill bias, human capital, redistribution, social contract, political economy.

JEL classification: D31, O33, J3, H10.

Roland Bénabou

Woodrow Wilson School and Department of Economics

Princeton University

Princeton, NJ 08544

rbenabou@princeton.edu

http://www.wws.princeton.edu/~rbenabou/

Introduction

The distribution of human capital and income lies at the center of a nexus of forces that shape a country's economic, institutional and technological structure. This chapter develops a unified model to analyze these interactions and their implications for growth, emphasizing in particular the mechanisms that allow different socioeconomic structures to perpetuate themselves, and those pushing toward convergence.[1] The analysis centers around five main questions.

1. *Why do countries at similar levels of development choose widely different social contracts?* Redistribution –through taxes and transfers, unemployment and health insurance, education finance and labor market regulation– displays remarkable variations even among countries with similar economic and political fundamentals. I thus ask what makes both European-type welfare states and US-type, more laissez-faire social contracts sustainable in the long run, together with their respective levels of inequality.[2] I then examine the efficiency and growth properties of these two regimes (which cannot be Pareto ranked) and ask what shocks might cause each one to unravel. The model also sheds light on the contrasting historical development paths of North and South America, and on the more recent experience of East Asia versus Latin America.

2. *How does skill-biased technical and organizational change impact the viability of redistributive institutions?* Over the last twenty-five years, most industrialized countries experienced a considerable rise in wage inequality.[3] This trend is generally attributed to three main factors: skill-biased technical change, international trade (which lies outside the scope of this chapter), and institutional change, such as the erosion of the minimum wage and the decline of unions. But minimum wages, labor market legislation and union power are endogenous outcomes, to the same extent as social insurance and education policy; and indeed, they evolved quite differently in Continental Europe or Canada and in the United States.[4] Analyzing redistributive institutions as a whole, I show how skill-biased technical change can cause the welfare state to unravel, and examine more generally how technological evolutions affect the set of social contracts that can be sustained in the long run.

The previous questions aim to explain differences in redistributive policies (together with their economic implications) and the role of technology in their evolution. The next two take the reverse perspective.

[1] The main channels through which inequality and redistributive institutions can in turn affect growth were exposited in Bénabou (1996).

[2] I shall limit my scope here to politico-economic persistence mechanisms that reflect differences in agents' economic interests and political power (Bénabou (2000), Saint Paul (2001), Hassler et al. (2003), Alesina, Glaeser and Sacerdote (2002)) rather than social norms (Lindbeck (1995)) or differences in beliefs about the mobility process and the determinants of individual income (Piketty (1995), Bénabou and Tirole (2002), Alesina and Angeletos (2003)).

[3] See, e.g. Autor, Katz and Krueger (1997) or Berman, Bound and Machin (1997).

[4] See, e.g., Freeman (1995), Fortin and Lemieux (1997), Lee (1999), or Acemoglu, Aghion and Violante (2001).

3. *What determines the types of technologies and organizational forms used by firms?* Production processes –and in particular their degree of skill bias– are themselves endogenous, adapting over time to the skills of the labor force.[5] I develop here a new and very tractable model of technology choice, based on the idea that firms tailor the *flexibility* of their production processes (substitutability between different labor inputs) to the distribution of human capital in the workforce. The main prediction is that the more heterogenous are workers' skill levels, the more flexible and wage-disequalizing the equilibrium technology will be. In a country like Japan, by contrast, production will involve much tighter complementarity between workers' tasks. Integrating this model with the previous analysis of human capital dynamics, I also show that firms' choices involve externalities that tend to result in excessive flexibility and a suboptimal growth rate, or even in self-sustaining technology-inequality traps.

4. *What types of societies and institutions are most conducive to the emergence of skill-biased technologies and organizational forms?* Through their influence on the distribution of human capital, public policies in the fiscal, labor market and especially educational arenas are important determinants of what innovations can be profitably developed and adopted; the same is true for immigration. One notes, for instance, that skill-biased technical change and reorganization occurred first, and to a greater extent, in the United States compared to Europe –and within the latter, more so in England than on the Continent. Combining the technology and policy components of the model, I show how a world-wide shift in the technological frontier leads to different evolutions of production processes and skill premia across countries with different social contracts.

Two extensive but essentially disconnected literatures have examined the economic determinants and consequences of redistributive policies on the one hand, those of biased technical change on the other.[6] Yet in reality both are endogenous, and jointly determined. The ability to conduct a unified analysis of human capital dynamics, technology and institutions is a novel and key feature of the framework developed in this chapter. It makes it possible to address important questions such as the second, fourth and especially fifth ones on the list:

5. *What "societal models" –joint configurations of technology, inequality, and policy– are feasible in the long run? In particular, how does the diffusion of technology affect nations' ability to maintain their own redistributive institutions and social structures?* Analyzing the case of two countries linked by the (endogenous) diffusion of their domestically developed technologies, I show how inequality

[5] See, e.g., Kremer and Maskin (1996), Acemoglu (1998), Kiley (1999), Lloyd-Ellis (1999), and Vindigni (1992). Relatedly, Grossman and Maggi (2000) show how the skill distribution matters for international specialization, and Legros and Newman (1996) how the wealth distribution affects the organization of firms.

[6] See the previously cited references, as well as the other ones given throughout the paper.

tends to be "exported" to the less heterogeneous one. This mechanism, in turn, generates spillovers between the social contracts of different nations, transmitting even purely political shocks and potentially triggering "chain reactions" that can cause major shifts towards a common, and generally inegalitarian, outcome.

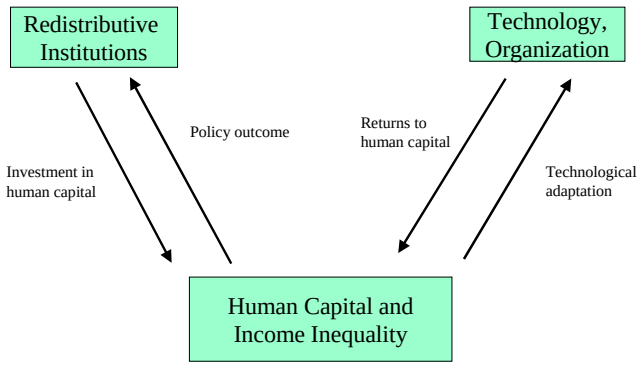


Figure 1: The links between inequality, technology and redistributive institutions

The chapter is organized in two main parts, corresponding respectively to the left- and right-hand sides of Figure 1.[7] The first of these two feedback loops centers on political-economy interactions. I thus present in Sections I and II a model of inequality, growth and redistributive policy in a context of imperfect credit and insurance markets (based on Bénabou (2000)). I first analyze how macro and distributional dynamics are affected by redistributive policies, then how the latter are themselves determined from the preferences and political power of different social classes. Finally, I identify the conditions under which a single or multiple politico-economic steady states arise.

The second and most novel part of the chapter incorporates the role of technology and its interactions with redistributive institutions. I first consider in Section III the impact of exogenous skill-biased technical change on inequality and the political equilibrium. I then study how technology responds to the composition of the labor force, through firms' choice of their degree of flexibility. In Section IV both sides of Figure I are brought together to analyze the long-run determination of institutions, technologies and the distribution of human capital. In Section V, finally, I show how technology diffusion leads to the "exporting" of inequality and international spillovers between social contracts. Section VI concludes. All proofs are gathered in the appendix.

[7] Each arrow on the diagram actually corresponds to a specific equation or proposition in the model. From left to right, these are (11), Proposition 3, (1) or later (28), and Proposition 8.

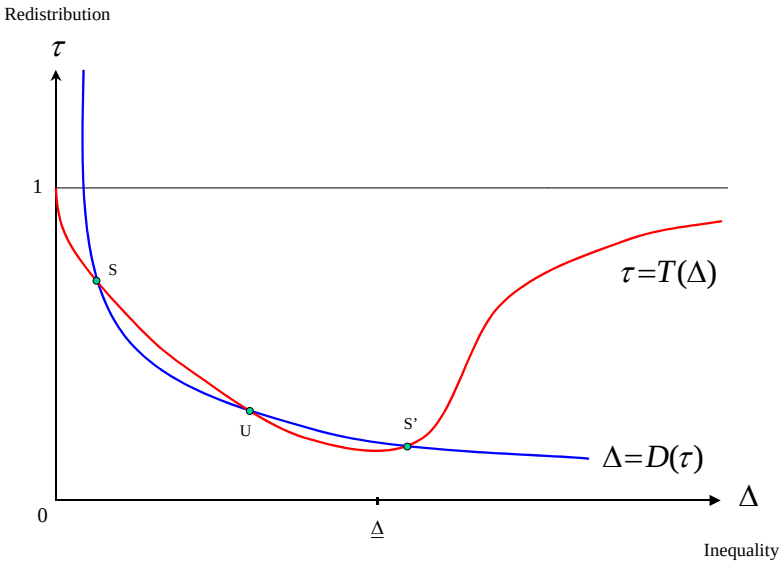


Figure 2: The two key relationships between inequality and redistribution.

I Inequality, Redistribution, and Growth

The model presented in this section (drawing on Bénabou (2000)) can be summarized by *two key relationships* between inequality and redistribution; both arise from imperfections in credit and insurance markets, and are illustrated on Figure 2.

The first locus summarizes the *political mechanism:* in each period, the equilibrium rate of redistribution chosen by voters is a U-shaped function $\tau = T(\Delta)$ of inequality in human capital, measured here as the variance of a lognormal distribution. The downward-sloping part of this curve, which is the crucial one, reflects a very general intuition: while asset market imperfections create a scope for efficient redistributive institutions (to provide social insurance and relax credit constraints), these institutions command much less support in an unequal society than in a relatively homogeneous one. Thus, starting from $\Delta = 0$, where there is unanimous support for the ex-ante efficient degree of redistribution, growing inequality increases the fraction of agents rich enough to lose from, and therefore oppose, all but relatively low levels of τ. The upward-sloping part of the curve, in contrast, is shaped by the standard skewness effect, which eventually dominates: rising numbers of poor will eventually impose more redistribution, well beyond the point where it ceases to be efficient.[8]

The second curve on Figure 2 represents the *accumulation mechanism:* since redistribution relaxes the credit constraints bearing on the poor's human capital investments, long-run inequality is a declining

[8] See, e.g., Alesina and Rodrik (1994) or Persson and Tabellini (1994) for models leading to such a positive slope. The empirical evidence (discussed at the end of this section) for both countries and US states provides little support for the standard view of a positive relationship between inequality and redistribution.

function $\Delta = T(\tau)$ of the rate of redistribution. When the two curves have several intersections, as illustrated on the figure, these correspond to *multiple politico-economic steady states* that are sustainable under the same fundamentals. One, with low inequality and high redistribution, corresponds to a European-type welfare state; the other, with the reverse configuration, to a US-type, more laissez-faire society.

In this and the next section I will derive the two loci from an explicit dynamic model, and identify the configurations of economic and political parameters under which alternative social models can coexist. In later sections I shall investigate how the two curves, and therefore the equilibrium set, are affected by exogenous technical change, then ultimately extend the analysis to the case where technology itself adapts endogenously to the distribution of skills in the population.

A *Production, Preferences and Policy*

The economy is populated by overlapping-generations families, $i \in [0,1]$. In generation t, adult i combines his human capital k_t^i with effort l_t^i to produce output, subject to a productivity shock z_t^i :

$$y_t^i = z_t^i \left(k_t^i\right)^\gamma \left(l_t^i\right)^\delta . \tag{1}$$

At this point the technology is exogenous and does not explicitly involve interactions among workers. Later on I will introduce a richer production structure, where agents with different skill levels perform complementary tasks and the degree of substitutability between them is optimally chosen by firms. The return to human capital γ and the mean of the productivity shocks z_t^i will then be endogenous functions of the current distribution of human capital. From the point of view of an individual worker-voter, however, this richer structure will retain an earnings function very similar to (1), so all the results obtained with this unconstrained reduced form will remain directly applicable.

Public policy or labor market institutions redistribute income through taxes and transfers, or a wage-equalization scheme, that transform each agent's gross earnings (or marginal revenue product) y_t^i into a disposable income $\hat{y}_t^i$, as specified further below. These resources finance both the adult's consumption, c_t^i, and his investment or educational bequest, e_t^i :

$$\hat{y}_t^i = c_t^i + e_t^i \tag{2}$$

$$k_{t+1}^i = \kappa\, \xi_{t+1}^i \left(k_t^i\right)^\alpha \left(e_t^i\right)^\beta, \tag{3}$$

where ξ_{t+1}^i represents the child's unpredictable ability, or simply luck, and $\alpha+\beta\gamma \leq 1$. There is thus no loan market for financing individual investments (e.g., children cannot be held responsible for the debts of their parents), and no insurance or securities market where the idiosyncratic risks z_t^i and ξ_{t+1}^i could be diversified

away.[9] Both shocks are i.i.d. and lognormal with mean one, and initial endowments are also lognormally distributed across families: thus $\ln z_t^i \sim \mathcal{N}(-v^2/2, v^2)$, $\ln \xi_t^i \sim \mathcal{N}(-w^2/2, w^2)$ and $\ln k_0^i \sim \mathcal{N}(m_0, \Delta_0^2)$.

Agents' preferences over their own consumption, effort, and child's human capital are defined recursively over their lifetime. Once he has learned his productivity z_t^i, agent i chooses his effort and consumption to maximize:

$$\ln V_t^i \equiv \max_{l_t^i,\, c_t^i} \left\{ (1-\rho)[\ln c_t^i - (l_t^i)^\eta] + \rho \ln E_t[k_{t+1}^i] \right\}. \tag{4}$$

The disutility of effort is measured by $\eta > 1$, which corresponds to an intertemporal elasticity of labor supply of $1/(\eta - 1)$. The discount factor ρ defines the relative weights of the adult's own felicity and of his bequest motive.[10]

At the beginning of period t, however, when evaluating and voting over redistributive policies, the agent does not yet know his lifetime productivity z_t^i. The resulting uncertainty over his *ex–post* utility level V_t^i is reflected in his *ex–ante* preferences, with a risk–aversion coefficient of a :

$$U_t^i \equiv \ln \left(E_t[(V_t^i)^{1-a} \mid k_t^i]^{1/(1-a)} \right), \tag{5}$$

This recursive specification allows a to parametrize the insurance value of redistributive policies, just as the labor supply elasticity $1/(\eta - 1)$ parametrizes the effort distortions.[11]

The redistributive policies over which agents vote are represented by simple, progressive schemes that map a market income y_t^i (marginal revenue product) into a disposable income $\hat{y}_t^i$, according to:

$$\hat{y}_t^i \equiv (y_t^i)^{1-\tau_t} (\tilde{y}_t)^{\tau_t}. \tag{6}$$

The break–even level $\tilde{y}_t$ is determined by the balanced-budget constraint, which requires that net transfers sum to zero. Thus, denoting per capita income by y_t, it must be that:

$$\int_0^1 (y_t^i)^{1-\tau_t} (\tilde{y}_t)^{\tau_t} \, di = y_t \,. \tag{7}$$

The elasticity τ_t measures the degree of *progressivity,* or equalization, of redistributive institutions.[12]

[9] The absence of any intertemporal trade is clearly an oversimplified (but quite common) representation of asset market incompleteness, making the model analytically tractable. Zhang (2004) extends a simplified version of the present model (with a zero-one policy variable and no political-economy mechanism) to allow for physical capital and financial bequests. He obtains similar results for the effects of inequality, plus new ones on convergence speeds to the steady-state.

[10] His (relative) risk–aversion with respect to the child's endowment k_{t+1}^i at that stage is normalized to zero, but this plays no role in any of the results. A dynastic specification of preferences (Bénabou (2002)) also leads to similar aggregate and distributional dynamics, but is less simple to work with.

[11] When $a \neq 1$ these recursive preferences are not time-separable (see, e.g., Kreps and Porteus (1979)), as risk-aversion differs from the inverse of the intertemporal elasticity of substitution in consumption, which by (4) remains fixed at one. This last assumption, common to many papers in the literature, helps make the model analytically solvable.

[12] When $\tau_t > 0$ the marginal rate rises with pretax income, and agents with average income are made better off: $\tilde{y}_t > y_t$. The elasticity of aftertax to pretax income is indeed the "right" measure of equalization: the posttax distribution induced by a fiscal scheme Lorenz–dominates the one induced by another (for all pretax distributions), if and only if the first scheme's elasticity is everywhere smaller (Fellman (1976)).

Three types of redistributive mechanisms can be considered here, being close to formally equivalent in this model. The first one, on which the exposition will generally focus, is that of fiscal policy, which equalizes disposable incomes through *taxes and transfers.* A second is *wage or earnings compression* through labor market institutions and policies favorable to workers with relatively low skills: minimum wage laws, union-friendly or right-to-strike regulations, firing costs, public sector pay and employment, etc.[13] The third one is *education finance,* where τ_t now applies only to human capital expenditures e_t^i, as opposed to all of income y_t^i. This may be achieved through a policy of school funding equalization across local communities, the presence of a centrally financed public-education system, or more generally by subsidizing differentially the education of rich and poor students.[14] Under either of the three above interpretations of τ_t, incentive compatibility requires that $\tau_t \leq 1$; on the other hand a regressive policy $\tau_t < 0$ cannot be ruled out a priori, and indeed one does observe such policies, typically in countries characterized by high inequality and a powerful ruling class.

B *Distributional Dynamics and Aggregate Growth*

Taking policy as parametrically given for the moment, I first consider the resulting economic decisions of individual agents, then the economy-wide dynamics of human capital and income.

Proposition 1 *Given a rate of redistribution τ_t, agents in generation t choose a common labor supply and savings rate: $l_t = \chi\,(1-\tau_t)^{1/\eta}$ and $e_t^i = s\,\hat{y}_t^i$, where $\chi^\eta \equiv (\delta/\eta)(1-\rho+\rho\beta)/(1-\rho)$ and $s \equiv \rho\beta/(1-\rho+\rho\beta)$.*

The fact that savings are unaffected is due to the imperfect-altruism assumption made regarding preferences.[15] Labor supply, on the other hand, declines in τ_t with an elasticity of $1/\eta$, and this single distortion will suffice to demonstrate how the efficiency costs and benefits of redistributive institutions shape the set of politico-economic equilibria.

Given Proposition 1, and substituting (6) into (3), the law of motion for human wealth is loglinear:

$$\begin{aligned} \ln k_{t+1}^i &= \ln \xi_{t+1}^i + \beta(1-\tau_t)\ln z_t^i + \ln\kappa + \beta \ln s \\ &\quad +(\alpha + \beta\gamma(1-\tau_t))\ln k_t^i + \beta\delta(1-\tau_t)\ln l_t + \beta\tau_t \ln \tilde{y}_t. \end{aligned} \tag{8}$$

[13] With the "autarkic" production function (1) the equivalence between the wage-income-equalization and the fiscal-redistribution interpretations of τ_t is immediate. It continues to hold when we move in Section III.B to a richer production structure with interacting agents.

[14] See Bénabou (2000) for this version of the model. Some of the formulas change slightly from those presented here for fiscal policy, but without affecting the qualitative nature of any of the results. There are, on the other hand, important quantitative differences between the growth and welfare implications of the two policies; see Bénabou (2002) and Sheshadri and Yuki (2004) for comparative analyzes. Previous models of redistribution centering on education finance include Becker (1964), Loury (1981), Glomm and Ravilkumar (1992), Saint-Paul and Verdier (1993), Bénabou (1996b) and Fernandez and Rogerson (1996). On the empirical side, see Krueger (2002) for a comprehensive summary and discussion of the evidence on targeted education and training policy interventions, from preschool to the college level.

[15] In Bénabou (2002) I develop and calibrate a version of the present model with dynastic preferences, where τ_t does affect the savings rate. On the other hand, agents are then able (and will indeed want) to use additional policy instruments, such as consumption taxes and investment subsidies, to alleviate this distortion.

This linearity reflects the absence of any non-convexities in the model, making clear that the multiplicity of equilibria will arise solely through the general-equilibrium feedback from the income distribution onto the political determination of τ_t.[16] These simple conditional dynamics also imply that human capital and income always remain lognormally distributed across agents:

$$\ln k_t^i \sim \mathcal{N}(m_t, \Delta_t^2), \tag{9}$$

$$\ln y_t^i \sim \mathcal{N}(\gamma m_t + \delta \ln l_t - v^2/2, \gamma^2 \Delta_t^2 + v^2), \tag{10}$$

where m_t and Δ_t^2 evolve according to two simple linear difference equations obtained by taking means and variances in (8), and given in the appendix. Since the growth of mean income y_t is of more direct economic interest than that of mean log-income m_t, I present here the equivalent characterization of the economy's dynamic path in terms of two linear difference equations in Δ_t^2 and $\ln y_t = m_t + \Delta_t^2/2$.

Proposition 2 *The distributions of human capital and income at time t are given by (9)-(10), where* $l_t = \chi(1-\tau_t)^{1/\eta}$. *The evolution of inequality across generations is governed by*

$$\Delta_{t+1}^2 = (\alpha + \beta\gamma(1-\tau_t))^2 \Delta_t^2 + \beta^2 (1-\tau_t)^2 v^2 + w^2, \tag{11}$$

and the growth rate of aggregate income by:

$$\ln(y_{t+1}/y_t) = \ln\tilde{\kappa} - (1-\alpha-\beta\gamma)\ln y_t + \delta(\ln l_{t+1} - \alpha \ln l_t) - \mathfrak{L}_v(\tau_t) v^2/2 - \mathfrak{L}_\Delta(\tau_t)\gamma^2\Delta_t^2/2, \tag{12}$$

where $\ln\tilde{\kappa} \equiv \gamma(\ln\kappa + \beta\ln s) - \gamma(1-\gamma)w^2/2$ *is a constant and*

$$\mathfrak{L}_v(\tau) \equiv \beta\gamma(1-\beta\gamma)(1-\tau)^2 \geq 0,$$

$$\mathfrak{L}_\Delta(\tau) \equiv \alpha + \beta\gamma(1-\tau)^2 - (\alpha + \beta\gamma(1-\tau))^2 \geq 0.$$

Equation (11) shows how inequality in the next generation stems from three sources: the varying abilities of children (w^2), shocks to family income (v^2), and differences in parental human capital (Δ_t^2), which matter both through family income and at-home transmission. Redistribution equalizes the disposable resources available to finance educational investments (but not social backgrounds), thus limiting both cross-sectional inequality and the *persistence* of family wealth, $\alpha + \beta\gamma(1-\tau_t)$; conversely, it increases *social mobility.*

Equation (12) makes apparent the growth losses from inequality due to credit constraints, and how redistribution's impact on growth involves a tradeoff between incentive and investment-allocation effects.[17]

[16] Or / and a feedback from the distribution onto the technology γ, once it is endogenized later on. By contrast, nearly all models in the literature that feature multiple equilibria rely on investment thresholds (e.g., Galor and Zeira (1993), Banerjee and Newman (1993)), indivisibilities in effort (Piketty (1997)), or non-homotheticity in preferences (e.g., Moav (2002)). For a discussion of indivisibilities, see also Mookerjee and Ray (2003).

[17] See Bénabou (1996b) for an overview of the literature on the relationship between inequality and growth, which is not the main focus of the present paper. In particular, inequality can also have positive effects on growth when there are non-

The effort distortion corresponds to the term in δ, which declines with parallel increases in τ_t and τ_{t+1}. The reallocation of human capital investments across differentially wealth-constrained agents is captured by the terms in $\mathfrak{L}_v(\tau_t)$ and $\mathfrak{L}_\Delta(\tau_t)$. When $\alpha = 0$ both are equal, and proportional to the concavity $\beta\gamma(1-\beta\gamma)$ of the common accumulation technology facing all families: differences in parental human capital and productivity shocks simply combine into variations in disposable income, $(1-\tau_t)^2\left(\gamma^2\Delta_t^2 + v^2\right)$, which credit constraints then translate into inefficient variations in investment, reducing overall growth proportionately. When $\alpha > 0$, however, disparate family backgrounds k_t^i represent complementary inputs that generate differential returns to investment, thus reducing the desirability of equalizing resources. Thus $\mathfrak{L}_\Delta(\tau)$ now differs from $\mathfrak{L}_v(\tau)$, and is minimized for $\tau = (1-\alpha-\beta\gamma)/(1-\beta\gamma)$, which decreases with α.

The term in $-\ln y_t$ in the growth equation, finally, reflects the standard convergence effect. It disappears under constant aggregate returns, namely when $\alpha + \beta\gamma = 1$, or when the constant κ in (3) is replaced by a knowledge spillover such as

$$\kappa_t \equiv \left(\int_0^i (k_t^i)^\gamma\right)^{(1-\alpha-\beta\gamma)/\gamma}. \tag{13}$$

This last variant yields an *endogenous-growth* version of the model, where all the predictions obtained with a constant κ in (12) now directly transpose from short-run growth and long-run per capita income to *long-term growth rates.*

Are the potential growth-enhancing effects of redistributive policies in the presence of credit constraints significant, or trivial compared to the standard deadweight losses? While the answer must ultimately come from empirical studies of specific policy programs or experiments, recent quantitative models suggest very important long-run effects, ranging from several percentage points of steady-state GDP to several percentage points of long-run growth, depending on the presence of accumulated factors, such as physical capital or knowledge spillovers, that complement individual human capital. Calibrating to US data a model with neither effort distortions nor complementarities, Fernandez and Rogerson (1998) find that complete school finance equalization raises long-run GDP by 3.2 %. In a model with both educational and financial bequests, Sheshadri and Yuki (2004) find that a mix of fiscal and educational redistribution that approximates current US policies raises long-run income by 13.5%, relative to laissez-faire. This more substantial impact primarily reflects the induced adjustment of physical capital, but it remains a level effect due to decreasing returns to the two types of capital together. In a dynastic-utility version of the present model with endogenous growth (Bénabou (2002)) I find that the growth-maximizing value for fiscal redistribution is $\tau_{fisc} = 21\%$, which corresponds to a share of redistributive transfers in GDP of 6%; in spite of reduced labor supply this raises the long-run growth rate by 0.5 percentage points. Under

convexities in either the investment technology (e.g., Galor and Zeira (1993)) or in preferences (e.g., Galor and Moav (1999)). For recent contributions to the empirical debate, see Forbes (2000) and Banerjee and Duflo (2003).

the alternative policy of progressive education finance, the growth-maximizing equalization rate for school expenditures is $\tau_{educ} = 62\%$, which raises long-run growth by 2.4 percentage points. In both cases, the efficient policy involves the top 30% of families subsidizing the bottom 70%, whether through the fiscal or the education system.

C Voter Preferences, Political Power, and Equilibrium Policy

I now turn to the determination of policy, which reflects both individual citizens' preferences and the allocation of power in the political system. In each generation, before the productivity shocks z_t^i are realized, agents vote on the rate of redistribution τ_t to which they will be subject; again, this could be through the fiscal system, labor market regulation, or education finance. Applying Propositions 1 and 2 to equations (4)-(5), an individual i's intertemporal welfare U_t^i can be computed from (5) as a function of the proposed policy τ_t, his endowment k_t^i, and the overall distribution of human capital (m_t, Δ_t), which is the system's state variable.[18] Defining the composite efficiency parameter

$$B \equiv a + \rho(1-a)(1-\beta) \geq 0, \tag{14}$$

whose interpretation is is given below, the resulting first-order condition for agent i's ideal tax rate takes the form:

$$\frac{\partial U_t^i}{\partial \tau_t} = (1-\rho+\rho\beta)\left[\gamma(m_t - \ln k_t^i) - \frac{\delta}{\eta}\left(\frac{\tau}{1-\tau}\right) + (1-\tau)(\gamma^2\Delta_t^2 + Bv^2)\right] = 0. \tag{15}$$

The first term inside the brackets, which disappears when summing across agents, reflects the basic redistributive conflict: since τ_t reallocates resources (spent on both consumption and children's education) from rich to poor households, the latter want it to be high, and the former, low. The next two terms represent the *aggregate welfare cost* and *aggregate welfare benefit* of a marginal increase in τ_t. First, there is the deadweight loss due to the distortion in effort: it is proportional to the labor supply elasticity $1/\eta$, and vanishes at $\tau = 0$. Second, the term $(1-\tau_t)(\gamma^2\Delta_t^2 + Bv^2)$, which is maximized for $\tau_t = 1$, embodies the (marginal) efficiency gains that arise from better insurance and the redistribution of resources towards more severely credit-constrained investments. Indeed it is clear from (14) that the composite parameter B multiplying the variance of adults' income shocks v^2 is monotonically related to both *risk–aversion* a and to the extent of *decreasing returns* in human-capital investment, $1-\beta$.[19] As to initial income inequality,

[18] See the appendix. Note that due to the model's overlapping–generations structure, voting involves no intertemporal strategic considerations.

[19] More specifically, under constant returns ($\beta = 1$) the term $(1-\rho+\rho\beta)Bv^2$ reduces to $a(1-\tau)v^2$, which is the insurance value of a marginal reduction in the lifetime resource risk $(1-\tau)^2v^2/2$ faced by agents. Conversely, for risk–neutral agents who care only about their offspring ($a = 0,\ \rho = 1$) that same term becomes $\beta(1-\beta)(1-\tau)v^2$, which is the gain in expected (and aggregate) human capital growth resulting from a marginal decrease in the variability of post-tax resources $(1-\tau)^2v^2/2$, given the concavity of the investment technology.

the term $\gamma^2\Delta_t^2$ reflects two motives for redistribution.[20] First, relaxing preexisting credit constraints tends to increase overall growth (see the last term in (12)), and therefore also average welfare. Second, with concave (logarithmic) utility functions, average welfare increases whenever individual consumptions (of c_t^i and k_{t+1}^i) are distributed more equally. Equivalently here, this captures the effect of skewness: given m_t, a higher Δ_t^2 implies a higher per capita income $\ln y_t = m_t + \Delta_t^2/2$, making redistribution more attractive for the median voter, and more generally at any given level of k_t^i.

From this analysis it easily follows that agent i's preferred tax rate, obtained as the unique solution $\tau_t^i < 1$ to the quadratic equation (15), decreases with his endowment k_t^i and increases with the ex-ante benefits from redistribution Bv^2. Similarly, $\left|\tau_t^i\right|$ decreases with $1/\eta$, as a more elastic labor supply magnifies the distortions that result from redistributive policies –whether progressive, $\tau > 0$, or regressive, $\tau < 0$.

I now turn from the preferences of different classes of voters to their political power or influence over the process that determines the actual τ_t. Even in advanced democracies, poor and less educated individuals have a lower propensity to register, turn out to vote and give political contributions, than better-off ones. For voting itself the tendency is relatively moderate, whereas for contributing to campaigns it is drastic. Even for political activities that are time- rather than money-intensive, such as writing to Congress, attending meetings, trying to convince others, etc., the propensity to participate rises sharply with income and education. These facts are documented for instance in Rosenstone and Hansen (1993), while Bartels (2002) provides a striking study of how they translate into disproportionate political influence. Studying the roll calls of US senators in three Congresses he finds that their votes are more responsive, by a factor ranging from 3 to 15, to the views of their constituents located the 75^{th} income percentile than to those of the 25^{th}; and again more responsive, by a factor of 2 to 3, to the views of the 99^{th} percentile than to those of the 75^{th}. In less developed countries there is also extensive vote-buying, clientelism, intimidation and the like, which are likely to result in even more bias.

To summarize this political influence of human and financial wealth in a simple manner I shall assume that *the pivotal voter is located at the* $100 \times p^*$ *-th percentile* of the distribution, where the critical level p^* can be any number in $[0, 1]$. A perfect democracy corresponds to $p^* = 1/2$, while an imperfect one where participation or influence rises with social status corresponds to $p^* > 1/2$.[21] Given that k_t^i is here log-normally distributed, an equivalent but more convenient measure of the political system's departure from the democratic ideal is

[20] See Bénabou (2000) for the exact decomposition.

[21] Since individual preferences are single-peaked and the preferred policy is monotonic in k_t^i, it is easy to show that such a critical p^* is a sufficient statistic for any *ordinal* weighing scheme where each agent's opinion is affected by a weight, or relative probability of voting, ω^i (with $\int_0^1 \omega^j \, dj = 1$), that increases with his *rank* in the distribution of human capital or income. Alternatively, political influence may depend on individuals' income *levels*. Thus, with ω^i proportional to $(y^i)^\lambda$ it can be shown that the pivotal voter has rank $p^* = \Phi(\lambda\Delta)$, so that λ in (16) is simply replaced by $\lambda\Delta$. As intuition suggests, this alternative formulation only reinforces the key result that efficient redistributions may decline with inequality, since it implies that the political system tends to becomes more biased towards the wealthy as inequality rises.

$$\lambda \equiv \Phi^{-1}(p^*), \tag{16}$$

where $\Phi(\cdot)$ denotes the c.d.f. of a standard normal. I shall refer to λ as the degree of *wealth bias* in the political system, and focus on the empirically relevant case where $\lambda > 0$.[22] Given the location of the pivotal voter, the policy outcome is simply obtained by setting $\ln k_t^i - m_t = \lambda \Delta_t$ in the first-order condition $\partial U_t^i / \partial \tau = 0$. This yields the quadratic equation:

$$\frac{1}{1-\tau_t} = \frac{1}{\lambda}\left[\frac{\gamma^2 \Delta_t^2 + Bv^2}{\gamma \Delta_t} - \frac{\tau_t}{\eta \gamma \Delta_t (1-\tau_t)^2}\right]. \tag{17}$$

When labor supply is inelastic ($1/\eta = 0$), it is immediately apparent that this equilibrium tax rate is U–shaped in Δ_t, and minimized where $\gamma^2 \Delta^2 = Bv^2$. This is true more generally.

Proposition 3 *The rate of redistribution $\tau_t = T(\Delta_t)$ chosen in generation t is such that:*

1) τ_t increases with the ex–ante efficiency gain from redistribution Bv^2, and decreases with the political influence of wealth, λ.

2) $|\tau_t|$ decreases with the elasticity of labor supply $1/\eta$.

3) τ_t is U-shaped with respect to inequality Δ_t. It starts at the ex–ante optimal rate $T(0) > 0$, declines to a minimum at some $\underline{\Delta} > 0$, then rises back towards $T(\infty) = 1$. The larger Bv^2, the wider the range $[0, \underline{\Delta})$ where $\partial \tau_t / \partial \Delta_t < 0$.

The first two results show that equilibrium policy depends on the costs and benefits of redistribution and on the allocation of political influence in a sensible manner. The third one confirms the key insight that efficient redistributions may *decrease* with inequality; more specifically, it yields the U-shaped function $\tau = T(\Delta)$ shown on Figure 2. The underlying intuition is simple, and very general: a) when distributional conflict $\gamma \Delta$ is small enough relative to the ex-ante efficiency gains Bv^2, there is widespread support for the redistributive policy, so its equilibrium level is high; b) as inequality rises, so does the proportion of agents rich enough to be net losers from the policy, who will *block* all but relatively low levels of τ_t; c) at still higher levels of inequality, the standard skewness effect eventually dominates: there are so many poor that they *impose* high redistribution, even when it is very inefficient.[23]

It is now well-recognized that the standard median-voter model's prediction of a positive effect of inequality on redistribution fails to explain the empirical patterns actually observed, both across countries (see, e.g., Perotti (1996), Bénabou (1996a, 2000), Alesina et al. (2002)) and within them (see Rodriguez

[22] Recent papers that aim to endogenously explain the allocation of political power in a country (corresponding here to the parameter λ) include Bourguignon and Verdier (2000), Pineda and Rodriguez (2000), Acemoglu and Robinson (2000), and Baland and Robinson (2003).

[23] A similar form of non-monotonicity (U-shape, or even declining throughout for λ high enough) is obtained with a Pareto distribution by Lee and Romer (1998).

(1999) for panel-data tests on US states). Among developed countries, in particular, the relationship is in fact negative (Pineda and Rodriguez (2000). The present framework explains how and when greater inequality will indeed *reduce* redistribution, or even result in *regressive* policies –both in the short run (Proposition 3) and in the long-run, when both are endogenous (Proposition 4 below). Furthermore, the distinctive *non–monotonic* relationship predicted by the model turns out to have empirical support: in tests using cross-country data, Figini (1999) finds in a significant U-shaped effect of income inequality on the shares of tax revenues and government expenditures in GDP; De Mello and Tiongson (2003) find a similar pattern for government transfers.

II Sustainable Social Contracts

A *Dynamics and Steady States*

The joint evolution of inequality and policy is described by the recursive dynamical system:

$$\left\{ \begin{array}{lcl} \tau_t & = & T(\Delta_t) \\ \Delta_{t+1} & = & \mathfrak{D}(\Delta_t, \tau_t) \end{array} \right. \tag{18}$$

where $T(\Delta_t)$ is given by Proposition 3 and $\mathfrak{D}(\Delta_t, \tau_t)$ by (11). Under a time–invariant policy, in particular, long–run inequality decreases with redistribution:

$$\Delta_\infty^2 = \frac{w^2 + \beta^2(1-\tau)^2 v^2}{1-(\alpha+\beta\gamma(1-\tau))^2} \equiv D^2(\tau). \tag{19}$$

A steady-state equilibrium is an intersection of this downward-sloping locus, $\Delta = D(\tau)$, with the U-shaped curve $\tau = T(\Delta)$, as illustrated in Figure 2. The following key proposition identifies the conditions under which multiple intersections occur.

Proposition 4 *Let $1-\alpha < 2\beta\gamma$. When the normalized efficiency gain B is below some critical value $\underline{B}$ there is a unique, stable, steady–state. When $B > \underline{B}$, on the other hand, there exist $\underline{\lambda}$ and $\bar{\lambda}$ with $0 < \underline{\lambda} < \bar{\lambda}$, such that:*

1) For each λ in $[\underline{\lambda}, \bar{\lambda}]$ there are (at least) two stable steady states.[24]

2) For $\lambda < \underline{\lambda}$ or $\lambda > \bar{\lambda}$ the steady–state is unique.

These results can shed light on a number of important issues and puzzles raised in the introduction.

First, they explain how countries with similar economic and political fundamentals can nonetheless sustain very different redistributive institutions, such as a *European-style welfare state* and a *US-style*

[24] See Bénabou (2000) for additional results on the number of stable steady-states ($n \leq 4$), including conditions ensuring that $n = 2$.

laissez-faire social contract. Notably, these two societies cannot be Pareto ranked. Recall also that τ_t can be equally interpreted as describing tax-and-transfer policy, labor market regulation, or (with some minor changes) education finance policy. Moreover, it is clear that the model's key mechanism makes these multiple dimensions of policy complementary, so that they will tend to covary positively across countries, as indeed they do empirically. A more egalitarian education system, for instance, tends to reduce income inequality, which in turn increases political support for fiscal redistribution or labor-earnings compression –and vice-versa. Summarizing a large collective research project on Sweden, Freeman (1995) emphasizes the presence of such complementarities, describing *"a highly interrelated welfare state and economy in which many parts fit together (be they subsidies, taxes, wage compression etc.)".*

Second, the two conditions required for multiplicity embody *very general intuitions* that are easily understood in the context of Figure 2. To start with, the ex-ante welfare benefits of redistribution must be high enough, relative to the costs.[25] Otherwise the T curve will be upward-sloping except over a very narrow initial range, and consequently have a unique intersection with the D curve; economically speaking, we would be close to the standard, complete-markets case. In addition, the political power of the wealthy must lie in some intermediate range, otherwise the T curve will lie too high or too low relative to the D curve, and again there will be a unique intersection, with high inequality and low redistribution, or vice-versa.

Third, while in the short-run the relationship is non-monotonic, there emerges in the long–run a *negative* correlation between inequality and redistribution, as indeed one observes between the United States and Europe, or among advanced countries in general (Pineda and Rodriguez (2000)).

Fourth, *history matters* in an important and plausible way: temporary shocks to the distribution of wealth (immigration, educational discrimination, demand shifts) as well as to the political system (slavery, voting rights restrictions) can permanently move society from one equilibrium to the other, or more generally have long-lasting effects on inequality, growth, and institutions. In particular, the model provides a formalization of Engerman and Sokoloff's (1998) thesis about the historical origins of South and North America's very different development paths, which they trace back to the former set of New World colonies having had much higher initial inequality Δ_0, and a much more concentrated power structure λ_0, than the latter.[26]

Finally, the model also shows that *different sources of inequality* have different effects on redistributive institutions –which, in particular, sheds doubt on the possibility of empirically estimating a catch-all relationship between inequality and redistribution, or inequality and growth. Indeed, one can show (provided

[25] The claim with respect to the benefits is clear from Proposition 4; with respect to the costs one can show, under additional technical assumptions, that the threshold $\underline{B}$ shifts up as the labor supply elasticity $1/\eta$ rises.

[26] This, in turn, was due to reasons linked to the technologies required for the different goods these colonies were producing –a point I shall come back to in Section III.A.

$1/\eta$ is not too large) that the threshold for multiplicity $\underline{B}$ is a decreasing function of the variance ratio v^2/w^2, with $\lim_{v/w\to 0}(\underline{B}) = +\infty$ and $\lim_{v/w\to+\infty}(\underline{B}) = 0$. Quite intuitively, income *uncertainty* interacts with the incompleteness in insurance and credit markets in generating ex ante efficiency gains from redistribution, as reflected by the term Bv^2 in (17). By contrast, a greater variance w^2 of the endowments that agents receive *prior* to choosing policy increases the distributional conflict between *identifiable losers and gainers* from the policy. Thus, whereas an increase in the variability of sectoral shocks (similar to v^2) will lead to an expansion of the welfare state, a surge in immigration that results in a greater heterogeneity of the population (similar to a rise in w^2) can easily lead to cutbacks, or even a large–scale dismantling. We shall observe similar effects when studying the political implications of skill-biased technical change.

B *Which Societies Grow Faster?*

As mentioned earlier, the steady states corresponding to different social contracts are not Pareto–rankable: rich enough agents always prefer a more laissez-faire society, while those who are poor enough always want more of a welfare state. One may still ask, however, how these two social models compare in terms of aggregate growth. This question is important first for its policy content, and second to know whether one should expect any empirical relationship between inequality and growth, when account is taken of the fact that *both* are endogenous. The answer hinges on the basic tradeoff, discussed earlier, between the distortions induced by redistribution and its beneficial effect on credit–constraints (magnified, in the long run, by the fact that it also reduces income inequality $\gamma\Delta_\infty$). This is made clear by the following results, which apply equally in the short and in the long run.[27]

Proposition 5 *Compared to a more laissez-faire alternative τ', a more redistributive social contract $\tau > \tau'$ is associated with lower inequality, and*

1) has higher growth when tax distortions are small $(1/\eta \approx 0)$ relative to those induced by credit constraints on the accumulation of human capital $(\beta\gamma < 1)$;

2) has lower growth when tax distortions are high $(1/\eta > 0)$ and the credit–constraint effect is weak $(\beta\gamma \approx 1)$.

The first scenario, of *"growth-enhancing redistributions",* seems most relevant for developing countries, where capital markets are less well-functioning, and for redistribution through public investments in human capital and health. One may contrast here the paths followed by East Asia and Latin America in those respects. The result may also help understand why regression estimates of the effects of social and educational transfers on growth are often significantly positive, or at least rarely significantly negative.

[27] See Section I.B for the simple correspondence between the stationary and the endogenous-growth versions of the model, where policy affects growth in the short and the long-run respectively.

The second, *Eurosclerosis"* scenario can account for why Europeans consistently choose more social insurance than Americans –at the cost of higher unemployment and slower growth –even though they are not necessarily more risk-averse. The intuition is that, in more homogenous societies, there is less erosion of the consensus over social insurance mechanisms which, ex–ante, would be valued enough to compensate for lesser growth prospects.[28]

Putting the two cases together, finally, Proposition 5 can also be related to the empirical findings of Barro (2000) that inequality tends to be negatively associated with subsequent growth in poor countries, but positively associated with it in richer ones. To the extent that poor countries are also those where credit markets are least developed, Proposition 5 predicts that inequality-reducing policies will give rise to just such a dichotomy.

III Technology and the Social Contract

I shall now extend the model to analyze how technology and redistributive institutions both affect inequality and respond to it, and consequently how they *influence each other* –as described on Figure 1. Of particular interest are the following questions. First, how does technical change impact the sustainability of welfare-state and laissez-faire social contracts? Second, what types of societies are likely to be leaders or early adopters in developing or implementing flexible, skill-biased technologies or organizational forms? More generally, how do the skill distribution among workers and the production side of the economy shape each other, through human capital investments and technology choices? Finally, what happens in the long run when technological and institutional factors evolve interdependently –within a country, and possibly even across countries?

A *Exogenous Technical Change and the Viability of the Welfare State*

I first examine here how technical or organizational change that increases the return to human capital affects redistributive institutions. This policy response represents an additional channel through which technological evolutions affect the income distribution, in addition to their direct impact via the wage structure.

Figure 3 illustrates the effects of an increase in γ, the coefficient on human capital in the production and earnings function (1). As will from now on be made explicit in the notation, this affects both of the key curves describing the inequality-redistribution nexus:

[28] For the specific case of unemployment insurance, Hassler et al. (1999) provide a complementary explanation, based on interactions with workers' specialization (or lack thereof) that can result in multiple equilibria.

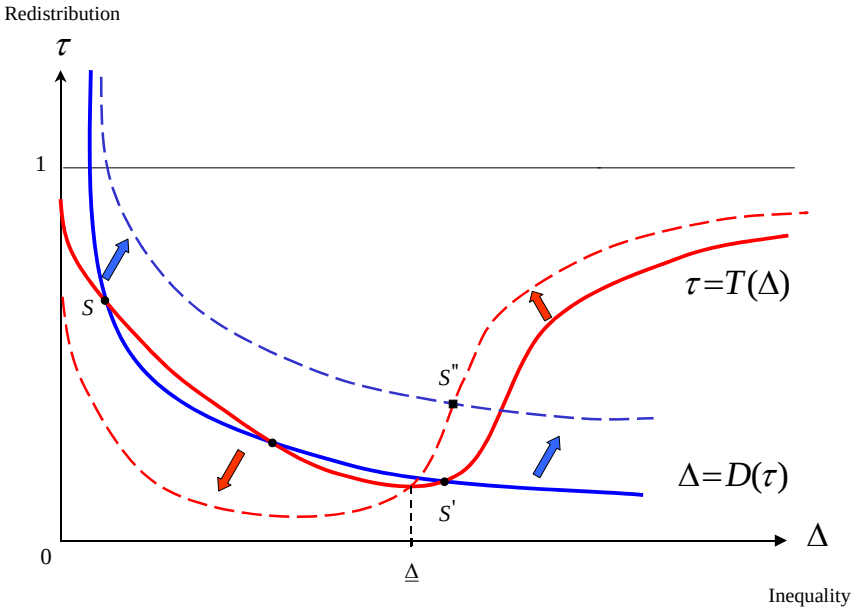


Figure 3: The effects of an increase in the returns to human capital, $\gamma = (\sigma - 1)/\sigma$.

i) The intergenerational-transmission locus $\Delta = D(\tau; \gamma)$ shifts up, and becomes less steep: for given human capital inequality Δ_t and policy τ there is more inequality in incomes $\gamma\Delta_t$, hence also in investments, and consequently more inequality of human capital (and of course income) in all subsequent periods.[29]

ii) The policy locus $\tau = T(\gamma\Delta)$ shifts down over $[0, \underline{\Delta})$, and up over $(\underline{\Delta}, +\infty)$: since what matters for the political outcome is *income* inequality $\gamma\Delta$ (see (17), an increase in γ for given Δ has the same U-shaped effect on redistribution as an increase in Δ for given γ –initially lowering τ, then raising it.

Figure 3 directly yields a local analysis of the more egalitarian, welfare-state equilibrium –and more generally, of any steady state that occurs along the declining portion of the T locus.[30]

Proposition 6 *Let (Δ, τ, γ) be such that (Δ, τ) is a stable steady state under the technology γ, with $\Delta < \underline{\Delta}(Bv^2; \gamma)$. A marginal increase in γ results in higher long run human capital and income inequality, as well as in less redistribution.*

[29] Recall that a worker's human capital reflects his individual ability, family background, and parental investment in education: $k_t^i = \kappa\, \xi_t^i\, (k_{t-1}^i)^\alpha\, (e_{t-1}^i)^\beta$. The kind of technical change considered here raises the return to all three components of k_t^i equally. In Galor and Tsiddon (1997) by contrast, major innovations raise the relative return to pure ability, while subsequent learning-by-doing innovations raise the relative return to inherited human capital. In Galor and Moav (2000) human capital is also sector-specific, and therefore eroded by new technologies, to an extent that decreases with individual ability. In these models technological innovations can thus raise as well as lower integenerational mobility.

[30] For steady-states that occur on the rising part, local comparative statics are ambiguous. Note, however, that in versions of the model where power inequality rises with income or human wealth inequality –meaning that λ increases with Δ (see footnote 21)– the declining portion of the locus is wider and the increasing portion reduced, making it easier to rule out such equilibria. For instance, if political power ω_i is proportional to $(y_i)^\lambda$ –e.g., "one dollar, one vote" for $\lambda = 1-$ then λ is simply replaced by $\lambda\Delta$ everywhere. As seen from (17), for $1/\eta = 0$ the $T(\gamma\Delta)$ curve is then decreasing throughout.

The policy response thus amplifies the direct effect of skill-biased technical progress on disposable incomes –and, over time, on the distributions of human capital and earnings. Figure 3 also suggests that it can have, in the long run, much more drastic consequences for redistributive institutions: starting from a situation with multiple steady states, an increase in γ tends to undermine the sustainability of the "Welfare State" equilibrium. Similarly, we shall see that starting from a configuration with a single "Welfare-State" it can make a second, "Laissez-Faire" equilibrium appear. Such a *global* analysis is potentially quite complicated, however, since in general there may be more than two stable equilibria, and some may also occur in the upward-sloping portion of the $\tau = T(\gamma\Delta)$ locus, where the policy response has a dampening rather than an amplifying effect on inequality. To demonstrate the most interesting insights, I shall therefore impose some simplifying assumptions. First, I restrict voters to a choice between only two policies:

• A generous *"Welfare State"* social contract, corresponding to a relatively high rate of redistribution $\bar{\tau} \in (0, 1)$;

• A more *"Laissez Faire"* social contract, corresponding to a relatively low rate of redistribution $\underline{\tau} \in (0, \bar{\tau})$.

Once again, τ can be interpreted as corresponding to either fiscal redistribution, wage compression through labor market regulation, or education finance progressivity. To further simplify the problem I abstract from labor supply distortions ($1/\eta = 0$) and assume that B is large enough that both potential steady states are always on the downward-sloping part of the $\tau = T(\Delta\gamma)$ curve, which is the one of most interest.[31]

Given an initial distribution of human capital Δ_t, the more redistributive policy $\tau_t = \bar{\tau}$ is adopted over $\tau_t = \underline{\tau}$ if $U_t^i(\bar{\tau}) > U_t^i(\underline{\tau})$ for at least a critical fraction $p^* \equiv \Phi(\lambda)$ of the population. Note from (15) that with $1/\eta = 0$, $\partial U_t^i/\partial\tau_t$ is linear in τ_t, so the preceding inequality evaluated at $\ln k_t^i = m_t + \lambda\Delta_t$ takes the form:

$$(\bar{\tau} - \underline{\tau})\left[\gamma\lambda\Delta_t + (1 - \underline{\tau})\left(\gamma^2\Delta_t^2 + Bv^2\right)\right] < (\bar{\tau} - \underline{\tau})^2\left(\gamma^2\Delta_t^2 + Bv^2\right)/2, \quad \text{or:}$$

$$\lambda < \left(1 - \frac{\bar{\tau} + \underline{\tau}}{2}\right)\left(\gamma\Delta_t + \frac{Bv^2}{\gamma\Delta_t}\right). \tag{20}$$

We first see that the *political influence of wealth* must not be too large, compared to the *aggregate welfare gain* from redistribution relative to laissez faire (net of the deadweight loss, which I am here abstracting from). Second, preexisting income inequality *raises the hurdle* that public policy must overcome, as the ex-ante benefit term Bv^2 is divided by $\gamma\Delta_t$. This effect impedes the adoption of more redistributive

[31] The required condition appears in Proposition 7. It is thus not inevitably the case that skill-biased technical progress leads to a retrenchment of redistributive institutions; the model allows for the reverse case, for steady-states that occur on the rising part of the T locus. The case on which I focus, however, appears to be the most relevant for recent trends, and in any case is the more robust, since: i) when multiple steady-states exist, there is always at least one the declining part; ii) in simple and plausible variants of the model, the T locus is decreasing throughout (see footnote 30).

institutions ($\tau = \bar{\tau}$) where they had not previously been in place, because of the greater divergence of interests that results over time from a more laissez-faire system ($\tau = \underline{\tau}$). Pushing in the other direction – namely, intensifying the demand for redistribution as inequality rises– are the effects of skewness and initial credit-constraints, reflected in the additive term $\gamma\Delta_t$. As a result of these offsetting forces, the right-hand side of (20) is U-shaped in $\gamma\Delta_t$. To focus on the long–run, let us now replace human capital inequality Δ_t with its asymptotic value under a technology γ and a constant policy τ –namely, by (11):

$$D(\tau,\gamma) \equiv \sqrt{\frac{w^2+\beta^2(1-\tau)^2v^2}{1-(\alpha+\beta\gamma(1-\tau))^2}}, \tag{21}$$

which is the long-run inequality in human capital resulting from a constant policy τ and technology γ. Given γ, the policy-inequality pair $(\bar{\tau}, D(\tau,\gamma))$ is thus a politico-economic steady state if:

$$\lambda < \left(1-\frac{\bar{\tau}+\underline{\tau}}{2}\right)\left(\gamma D(\tau,\gamma)+\frac{Bv^2}{\gamma D(\tau,\gamma)}\right) \equiv \bar{\lambda}(\gamma;B). \tag{22}$$

Conversely, the laissez-faire configuration $(\underline{\tau}, D(\underline{\tau},\gamma))$ is a politico-economic steady state given γ if:

$$\lambda > \left(1-\frac{\bar{\tau}+\underline{\tau}}{2}\right)\left(\gamma D(\underline{\tau},\gamma)+\frac{Bv^2}{\gamma D(\underline{\tau},\gamma)}\right) \equiv \underline{\lambda}(\gamma;B). \tag{23}$$

The two regimes coexist if and only if $\underline{\lambda}(\gamma;B) < \bar{\lambda}(\gamma;B)$, or:

$$\frac{\bar{\lambda}(\gamma;B)-\underline{\lambda}(\gamma;B)}{\gamma D(\underline{\tau},\gamma)-\gamma D(\bar{\tau},\gamma)} = \left(1-\frac{\bar{\tau}+\underline{\tau}}{2}\right)\left(\frac{Bv^2}{\gamma^2 D(\bar{\tau},\gamma)D(\underline{\tau},\gamma)}-1\right). \tag{24}$$

We thus obtain here the analogue, for a discrete policy choice, of Proposition 4: multiplicity requires that B be large enough compared to income inequality (and, in general, to $1/\eta$),

$$B > \left(\gamma^2/v^2\right)\cdot D(\bar{\tau},\gamma)\cdot D(\underline{\tau},\gamma) \equiv \underline{B}(\gamma), \tag{25}$$

and that the wealth bias λ be neither too high nor too low, given the technology $\gamma : \lambda \in \left[\underline{\lambda}, \bar{\lambda}\right]$, defined by (22)-(23).[32] Now, furthermore, we shall see that (under appropriate conditions) the *skill bias* γ must also be neither too high nor too low, given λ. This result is illustrated in Figure 4.

Proposition 7 *Let $1/\eta = 0$ and $Bv^2 > \gamma_{\max}\cdot D(\underline{\tau},\gamma_{\max})$, where $\gamma_{\max} \equiv (1-\alpha)/\beta$. There exist two skill-bias thresholds $\underline{\gamma}(\lambda;B) < \bar{\gamma}(\lambda;B)$, both decreasing in λ and increasing in B, such that:*

[32] Note also that as B increases both $\underline{\lambda}$ and $\bar{\lambda}$ rise, but (24) shows that the interval $[\underline{\lambda},\bar{\lambda}]$ widens. When (25) does not hold, on the other hand, we have $\bar{\lambda} < \underline{\lambda}$. For $\lambda \notin [\bar{\lambda},\underline{\lambda}]$ there is a unique steady–state, but for $\lambda \in [\bar{\lambda},\underline{\lambda}]$ the economy can instead be shown to cycle between the two regimes, as in Gradstein and Justman (1997). This feature reflects the restriction of policy to a binary choice.

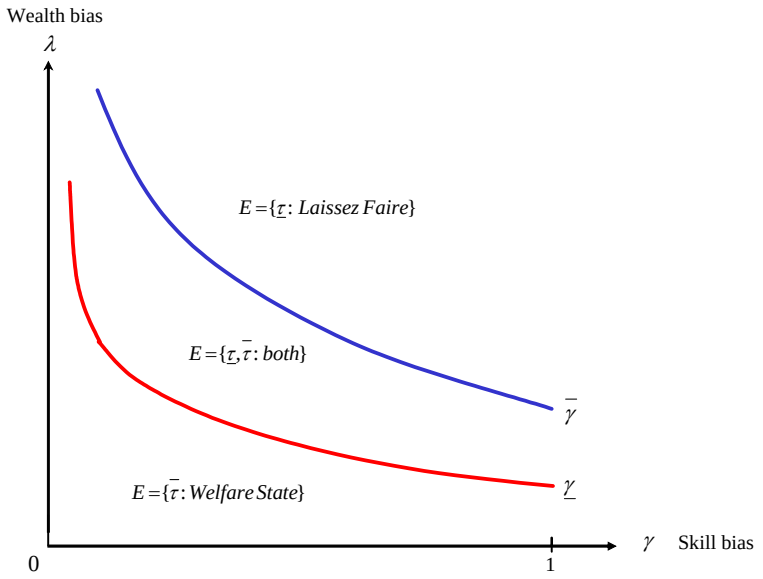


Figure 4: Technology, political influence, and the social contract. E denotes the set of stable steady-states.

i) for $\gamma < \underline{\gamma}(\lambda; B)$, *the unique steady state corresponds to the welfare-state outcome* $(\bar{\tau}, D(\bar{\tau}, \gamma))$;
ii) for $\gamma \in \left[\underline{\gamma}(\lambda; B), \bar{\gamma}(\lambda; B)\right]$, *both* $(\bar{\tau}, D(\bar{\tau}, \gamma))$ *and* $(\underline{\tau}, D(\underline{\tau}, \gamma))$ *are stable steady states;*
iii) for $\gamma \in [\bar{\gamma}(\lambda; B), \gamma_{\max}]$, *the unique steady state is laissez-faire,* $(\underline{\tau}, D(\underline{\tau}, \gamma))$.

These results have a number of important implications.

First, they confirm that *the Welfare State becomes unsustainable* when technology becomes too skill-biased; and, conversely, that multiple social contracts can coexist only when γ is in some intermediate range.[33] We see here again at work the general insight that sources of heterogeneity that are predictable on the basis of of initial endowments –a greater variance of abilities, w^2, as discussed earlier, or greater skill bias γ, as here– push equilibrium institutions towards less redistribution.

Second, Proposition 7 also reveals interesting interactions between *the production and political "technologies".* As seen on Figure 4, in a country with relatively little wealth bias the welfare state is –for better of for worse– much more "immune" to skill-biased technical change than in one where λ is high. Similarly, a given change in the political system will have very different effects on redistributive institutions, depending on how skill-biased the technology is. Finally, the "surest way" to set out on a course of persistently high inequality and inefficiently regressive (or insufficiently progressive) institutions is to start out with *both* a production structure that generates high wage inequality, and a political system marked by a high

[33] Hassler et al. (2003) also show that the "welfare-state" equilibrium in their model no longer exists above a certain level of skill bias. The mechanism is quite different, however: it is the *anticipation* of a higher skill premium that causes more agents to invest in education –to the point where, ex-post, a majority of them end up with high incomes (the distribution is negatively skewed), and therefore oppose redistribution.

degree of bias. As demonstrated by Engerman and Sokoloff (1995), such were the initial conditions found in the plantation-based and natural-resource based colonies of Central and South America in the 16^{th} and 17^{th}centuries –in contrast to those of North America, where agriculture was not subject to significant increasing returns to scale, and initial institutions were much less oligarchic.

Third, our result can also be related to that of Acemoglu, Aghion and Violante (2001), who show that skill-biased technical progress may cause a decline in unionization. While their model is quite different, it shares the two key features emphasized in previous sections. First, relatively rich agents –namely skilled workers– are pivotal, in the sense that it is their willingness to leave or avoid the unionized sector that limits the extent of wage compression. Second, in making this mobility decision –voting with their feet– they trade off redistributive losses (unions redistribute towards unskilled workers, who are a majority in the unionized sector) against ex-ante efficiency benefits: unions provide insurance through wage-sharing and / or a safeguard against the "holdup" by firms of workers' specific human capital investments; even when they play no such role, leaving the unionized sector involves mobility costs. Consequently, when skill-biased technical change makes the interests of the two classes of workers too divergent, redistributive institutions –here, union participation– will decline. Moreover, this can happen inefficiently.[34]

B *Skills, Technology, and Income Inequality*

I now turn to the reverse mechanism and examine how inequality itself *feeds back* onto the nature of technical change, making γ endogenous. Recognizing that individuals do not produce in isolation, I model production interactions with a simple specialization structure where workers perform complementary tasks.[35] Final output is produced by competitive firms, using a continuum of differentiated intermediate inputs:

$$y_t = A_t \cdot \left(\int_0^\infty z_t(s) \cdot x_t(s)^{\frac{\sigma-1}{\sigma}} \, d\, s \right)^{\frac{\sigma}{\sigma-1}} , \ \sigma \geq 1, \tag{26}$$

where $x_t(s)$ denotes the quantity of input s, $z_t(s)$ an i.i.d. sectoral shock, and A_t a TFP parameter. Workers specialize in a single good, which they produce using their human capital and labor. Since they face downward-sloping demand curves each selects a different task, $s(i) = i$, and produces $x_t^i = k_t^i \, l_t^i$ units, where l_t^i is endogenously chosen. The unit price for his output is thus:

$$p_t^i = A_t^{\frac{\sigma-1}{\sigma}} \cdot z_t^i \cdot (k_t^i \, l_t^i / y_t)^{-\frac{1}{\sigma}} \ . \tag{27}$$

The corresponding hourly wages are $\omega_t^i = p_t^i \, k_t^i$, and the resulting incomes

[34] Relatedly, note from Figure 4 that a minor change in γ can trigger a significant decline in redistribution from $\bar{\tau}$ to $\underline{\tau}$, and recall from Proposition 5 that the latter can easily lead to lower aggregate growth. The same is clearly true for average welfare, e.g. when $1/\eta = 0$.

[35] Building on those in Bénabou (1996) and Tamura (1992), themselves based on Romer (1987).

$$y_t^i = \omega_t^i \, l_t = z_t^i \cdot \left(k_t^i \, l_t^i\right)^{\frac{\sigma-1}{\sigma}} \times A_t^{\frac{\sigma-1}{\sigma}} (y_t)^{\frac{1}{\sigma}} \equiv \tilde{A}_t \cdot z_t^i \cdot \left(k_t^i\right)^{\gamma} (l_t^i)^{\delta}. \tag{28}$$

This earnings function is exactly the same as in previous sections (see (1)), with

$$\gamma = \delta \equiv \frac{\sigma - 1}{\sigma}, \tag{29}$$

except for the extra TFP factor $\tilde{A}_t \equiv A_t^{\frac{\sigma-1}{\sigma}} (y_t)^{\frac{1}{\sigma}}$, which acts as a shift in the mean of the productivity shocks z_t^i. While $\tilde{A}_t$ varies endogenously with the economy's state variables (m_t, Δ_t^2), individual workers and voters take it as given in their decisions over (l_t^i, c_t^i) and their votes over τ_t.[36] Consequently, the entire analysis of earlier sections still applies, with the simple substitution of $\tilde{A}_t \cdot z_t^i$ wherever z_t^i previously appeared. Conditional on γ, distributional dynamics and the political equilibrium thus remain essentially unchanged, and so do the corresponding $\Delta = D(\tau, \gamma)$ and $\tau = T(\gamma\Delta)$ loci.

I now consider firms. Recall that in equilibrium all workers supply the same effort $l_t^i = l_t$ and the distribution of human capital remains lognormal, $\ln k_t^i \sim \mathcal{N}(m_t, \Delta_t^2)$. The output of a representative firm is thus:

$$y_t = A_t \cdot l_t \cdot \left(\int_0^1 \left(k_t^i\right)^{\frac{\sigma-1}{\sigma}} \, di \right)^{\frac{\sigma}{\sigma-1}} = A_t \cdot l_t \cdot e^{-\Delta_t^2/2\sigma} \cdot \left(\int_0^1 k_t^i \, di \right). \tag{30}$$

Keeping average human capital constant, the loss $e^{-\Delta_t^2/2\sigma}$ makes apparent the productivity costs imposed by (excessive) *heterogeneity of the labor force:* poorly educated, insufficiently skilled production and clerical workers drag down the productivity of engineers, managers, scientists, etc. We also see that a production technology with greater substitutability between the tasks performed by different types of workers reduces these costs of skill disparities (Bénabou (1996), Grossman and Maggi (2000)). Indeed, this greater *flexibility* allows firms to more easily substitute towards the more productive workers, and conversely reduce their dependence on low-skill labor. This may be achieved by internal retooling, reorganization, or by outsourcing certain activities to competitive subcontractors.[37] One can also think of a higher σ as a more discriminating search technology, resulting in more assortative matching between workers –that is, in a more *segregated* production structure (Kremer and Maskin (1996), (2003)).[38]

Naturally, production processes with less complementarity between workers of different skills result in greater inequality of wages and incomes, as they have the effect of *uncoupling* their marginal products:

[36] Note again the role of the overlapping-generations structure with "imperfect" altruism in simplifying the voting problem. Observe also that τ_t can now, as claimed earlier, be interpreted as the extent of *wage income compression,* i.e. the degree of progressivity in the mapping (defined by (6)) from workers' true marginal revenue products y_t^i (given by (28)) to the labor earning they actually receive, $\hat{y}_t^i$.

[37] For evidence on organizational change, see for instance Caroli and Van Reenen (1999).

[38] When labor supply is endogenous, $1/\eta > 0$, a higher σ also induces workers to increase their labor supply, as they face a less elastic demand curve: by Proposition 1, $l_t = \chi (1 - \tau_t)^{1/\eta}$, with now $\gamma = (\sigma - 1)/\sigma$. This effect is independent of any issues of skill heterogeneity or wage inequality, however.

$$\text{Var}\left[\ln y_t^i\right] = \left(\frac{\sigma_t - 1}{\sigma_t}\right)^2 \Delta_t^2 = \gamma_t^2 \Delta_t^2. \tag{31}$$

C *Technological Choice and Endogenous Flexibility*

More flexible technologies and production processes require costly investments or reorganizations. Moreover, their benefits to an individual firm are endogenous even in the short run (i.e., given the skill composition of the labor force), as they depend on the decisions of other firms, which affect the wage structure.

I therefore now model firms' choices of technology or organizational form, proposing a new and very simple formulation that highlights the roles of heterogeneity and flexibility. In every period, firms have access to a menu of potential technologies with *different elasticities of substitution* $\sigma \in [1, +\infty)$ and associated costs $c(\sigma)$; the latter result in a TFP factor $A(\sigma) = e^{-c(\sigma)}$, with $c' > 0$ and $c'' > 0$.[39] Given the distribution of workers' human capital $\ln k_t^i \sim \mathcal{N}(m, \Delta_t^2)$ and the technology σ_t used by its competitors, each firm chooses its own technology $\hat{\sigma}$ as a best response. This results in a marginal cost of

$$A(\hat{\sigma})^{-1} \left(\int_0^1 \left(z_t^i\right)^{\hat{\sigma}} \left(p_t^i\right)^{1-\hat{\sigma}} di \right)^{\frac{1}{1-\hat{\sigma}}}. \tag{32}$$

Substituting from (27) for the equilibrium input prices p_t^i, and normalizing by the other firms' marginal cost (see the proof of Proposition 8), the firm's relative marginal cost is equal to:

$$mc(\hat{\sigma}|\sigma_t) = \left(\frac{A(\sigma_t)}{A(\hat{\sigma})}\right) \cdot \left(\int_0^1 \left(k_t^i\right)^{\frac{1-\hat{\sigma}}{\sigma_t}} di \right)^{\frac{1}{1-\hat{\sigma}}} \cdot \left(\int_0^1 \left(k_t^i\right)^{\frac{1-\sigma_t}{\sigma_t}} di \right)^{\frac{-1}{1-\sigma_t}},$$

or:

$$mc(\hat{\sigma}|\sigma_t) = \exp\left[c(\hat{\sigma}) - c(\sigma_t) + \frac{\Delta_t^2}{2} \left(\frac{\sigma_t - \hat{\sigma}}{\sigma_t^2}\right) \right]. \tag{33}$$

The first-order condition for this convex minimization problem is

$$c'(\hat{\sigma}) = \frac{\Delta_t^2}{2\sigma_t^2}. \tag{34}$$

Intuitively, the marginal benefit of flexibility rises with the variability of skills in the labor force, but decreases with the degree to which other firms choose technologies that allow them to more easily substitute toward better workers, since in doing so they drive up the skill premium.

[39] I thus abstract here from the intertemporal (investment) aspects of innovation that would be part of a more complete (but also more complicated) model of technological change; see, e.g., Acemoglu (1998), Kiley (1999), Lloyd-Ellis (1999), or Aghion (2002).

Proposition 8 *There is a unique symmetric equilibrium in technology choice. The more heterogenous the workforce, the more flexible and skill-biased the technology used by firms:* $\sigma_t = \sigma^*(\Delta_t)$ *is the solution to* $c'(\sigma^*) = \Delta^2/2(\sigma^*)^2$, *with* $0 < \partial \ln \sigma^*/\partial \ln \Delta < 1$.

This result has several interesting implications.

A first one is the *magnification of wage inequality:* the return to human capital $\partial \ln \omega_t^i/\partial \ln k_t^i = (\sigma_t - 1)/\sigma_t$ is higher where the labor force is more heterogenous, further amplifying wage differentials across educational levels. This simple prediction could be tested empirically across countries and / or time periods.[40]

A second implication is the potential for *"immiserizing technological choices".* Proposition 8 states that σ increases with Δ; conversely, because of credit constraints, human capital heterogeneity itself rises over time with $\gamma = (\sigma - 1)/\sigma$, and in the long-run $\Delta = D(\tau, \gamma)$, which is increasing in γ. Could these two mechanisms reinforce each other to the point of resulting in multiple steady states *even under a fixed policy* –whether activist or laissez-faire– and even though, once again, there are no non-convexities in the model? The idea is that a high degree of skill bias results in very low wages for unskilled workers, severely limiting the extent to which they can invest in human capital (for themselves or their children). This, in turn, leads firms to again choose a very flexible, skill-biased technology in the next period, and so on. Conversely, a less skilled-biased technology and a less dispersed distribution of human wealth could be self-sustaining. To examine this possibility, note first that:

$$\frac{\partial \ln \sigma^*}{\partial \ln \Delta} = \left(1 + \frac{1}{2}\frac{c''(\sigma_t)}{c'(\sigma_t)}\right)^{-1} < 1 \tag{35}$$

by Proposition 8, while (21) yields

$$\frac{\partial \ln D(\tau, \gamma)}{\partial \ln \sigma} = \frac{\beta(1-\gamma)(1-\tau)(\alpha + \beta\gamma(1-\tau))}{1 - (\alpha + \beta\gamma(1-\tau))^2} \tag{36}$$

where, as usual, $\gamma = (\sigma - 1)/\sigma$. If the product of these two derivatives is everywhere less than 1, there is a unique equilibrium. If it exceeds 1 for some value of σ, on the other hand, there may be multiplicity. It is easily verified that $\partial D(\tau, \gamma)/\partial \ln \sigma < 1$ if and only if

$$(\alpha + \beta\gamma(1-\tau))(\alpha + \beta(1-\tau)) < 1. \tag{37}$$

The first term is always less than one (or else inequality explodes; moreover, this can never occur when τ is

[40] Kremer and Maskin (1996) present evidence for a related intervening mechanism (similar to $\partial\sigma^*/\partial\Delta > 0$ in this model), although not for how educational returns and wage inequality are ultimately affected. They show that in US states characterized by greater human capital inequality, there is more segregation of workers by skills (the ratio of within- to between-plant skill dispersion is lower).

endogenously chosen), but the second need not be, especially if $\tau < 0$. We can thus conclude that the kind of *"technology-inequality trap"* described above becomes a real possibility under *regressive or insufficiently progressive* policies. In particular, education systems that result in significant resource disparities between students, such as private financing or local (property-tax based) school funding as in the United States, are fertile ground for the joint emergence of highly skill-biased production processes and a persistently skewed skill distribution. Furthermore, as we shall see below, endogenizing τ only increases the likelihood of such outcomes, since the degree of redistribution tends to fall with inequality.

A third point is that even under the less extreme conditions where no such trap exists, firms' decisions involve a *dynamic externality* that tends to result in *excessively skill-biased* or flexible technologies. Indeed, each takes the distribution of skills it faces as given but neglects the effects of its own flexibility on workers' human capital investments, and therefore on subsequent distributions. More specifically, while a marginal change in σ_t has only second-order effects on the current production costs faced by firms, it has three first-order effects on growth.[41] First, a lower σ_t would reduce current income inequality $\gamma_t \Delta_t$, which is growth-enhancing given the presence of credit constraints. This would in turn lower the skill disparities Δ_{t+k} that firms will face in the future, as well as the costs $c\left(\sigma^*(\Delta_{t+k})\right)$ they will bear to adapt to this heterogeneity. Although $\gamma_t = (\sigma_t - 1)/\sigma_t$ also affects in a somewhat complex way the concavity of educational investment (where it interacts with α, β and τ_t), it is easy to identify cases where growth *in every period* would be higher if firms collectively chose less skill-biased technologies.

For instance, let $\alpha = 0$, $\beta = 1$, and $1/\eta = 0$ (inelastic labor supply), and fix any constant policy τ; the interactions of technology choice and policy decisions will be examined in the next section. In the resulting steady state, the degree of flexibility and the dispersion in skills are given by the two equations $\sigma_\infty = \sigma^*(\Delta_\infty)$ and $\Delta_\infty = D(\tau, \gamma_\infty)$, where $\gamma_\infty \equiv (\sigma_\infty - 1)/\sigma_\infty$.[42] The corresponding asymptotic growth rate is computed in the appendix, and equals:

$$g_\infty = \ln \kappa + \ln s - c(\sigma_\infty) - \frac{D(\tau, \gamma_\infty)^2}{2\sigma_\infty}. \tag{38}$$

A marginal reduction in σ from its equilibrium value, if it were permanently implemented by all firms, would then increase steady-state growth, since:

$$\left.\frac{\partial g_\infty}{\partial \sigma}\right|_{\sigma=\sigma_\infty} = -c'\left(\sigma_\infty\right) + \frac{\Delta_\infty^2}{2\sigma_\infty^2} - \frac{1}{2\sigma_\infty} \cdot \left.\frac{\partial D^2(\tau, \gamma_\infty)}{\partial \sigma}\right|_{\sigma=\sigma_\infty} = -\frac{1}{2\sigma_\infty^3} \cdot \frac{\partial D^2(\tau, \gamma_\infty)}{\partial \gamma} < 0. \tag{39}$$

In this expression the first two terms cancel out by the first-order condition (34), while the last one reflects

[41] As explained in footnote 38, when $1/\eta > 0$ a higher σ_t also raises the return to labor supply $\delta_t = (\sigma_t - 1)/\sigma_t$, inducing all agents to work more.

[42] I assume here that (37) holds, so that this steady-state is unique (given τ), although this is inessential to the argument.

the dynamic externality. The above result holds more generally for any equilibrium path that is either near the steady state, or such that σ_t converges to its long-run value from above (see the appendix).

Inefficient choices of technology or firm organization arise in a number of models where market imperfections create an excessive role for the distribution of financial or human wealth to shape the structure of production, with the result of exacerbating inequality and making it more persistent. In Banerjee and Newman (1993) and Newman and Legros (1998), for instance, the moral-hazard problem affecting entrepreneurship combines with an unequal wealth distribution in forcing too many agents to work for low wages in large firms, rather than setting up their own. In Vindigni (2002) an extreme example of the technology trap studied above occurs, as firms' decisions (choosing the arrival rate of exogenously skill-biased innovations) can permanently confine some dynasties of workers below the fixed income threshold required to invest in human capital.[43] In Grossman (2004), a high variance of human capital in the labor force increases the incentives of the most skilled agents to work in sectors where individual productivity is observable, rather than in those where output is team-determined; because they fail to internalize the spillovers they would have on team productivity, the resulting occupational segregation is inefficiently high.

IV Endogenous Institutions and Endogenous Technology

Combining the main mechanisms analyzed in previous sections yields a model where the distribution of human capital, the technologies used by firms and the policy implemented by the state are all endogenous –as they are in reality. The dynamical system governing the economy's evolution remains recursive:

$$\left\{ \begin{array}{rcl} \gamma_t & = & \Gamma(\Delta_t) \\ \tau_t & = & T(\Delta_t \gamma_t) \\ \Delta_{t+1} & = & \mathcal{D}(\Delta_t, \tau_t; \gamma_t) \end{array} \right. , \tag{40}$$

where $\Gamma(\Delta) \equiv (\sigma^*(\Delta) - 1)/\sigma^*(\Delta)$ represents the technology outcome given by Proposition 8, $T(\gamma\Delta)$ the policy outcome given by Proposition 3, and $\mathcal{D}(\Delta, \tau, \gamma)$ the transmission of human capital inequality given in Proposition 2. The resulting aggregate growth rate, $\ln(y_{t+1}/y_t) = g(\tau_t, \Delta_t, \gamma_t)$, follows from Proposition 2. Finally, steady states are solutions to the fixed-point equation

$$\Delta = \mathcal{D}(\Delta, T(\Delta; \Gamma(\Delta)), \Gamma(\Delta)). \tag{41}$$

[43] A more benign form of multiplicity (with greater wage inequality now going together with more, rather than less, total human capital) occurs in Acemoglu (1998). In his model, a relative abundance of skilled workers makes it more profitable for firms to develop skill-biased technologies; this then raises the wage premium, encouraging more workers to become skilled.

This structure makes clear the presence of important *multiplier effects:* a transitory shock affecting inequality (e.g., more idiosyncratic uncertainty v^2) or the political system (a higher λ) will be amplified through technology decisions, the policy choice, and the intergenerational transmission mechanism, and may thus have considerable long-term consequences.[44] Most importantly, in accounting for changes in inequality *one can no longer treat technological and institutional factors as separate,* competing explanations: both are jointly determined, and complementary. The model thus shows how, in the words of Freeman (1995), one needs to think of *"the Welfare State as a system".*

To demonstrate these points I shall assume from here on a piecewise-linear technological frontier. Flexibility is free up to σ_L, then has a marginal cost of $M > 0$, up to a maximum level $\sigma_H > \sigma_L$:

$$c(\sigma) = \begin{cases} 0 & \text{for} \quad \sigma < \sigma_L \\ M(\sigma - \sigma_L) & \text{for} \quad \sigma \in [\sigma_L, \sigma_H] \\ +\infty & \text{for} \quad \sigma > \sigma_H \end{cases} . \tag{42}$$

I will denote $\gamma_i = (\sigma_i - 1)/\sigma_i$, $i \in \{L, H\}$. The analogue of Proposition 8 in this case is very simple, as the first order condition in a symmetric equilibrium involves the comparison:

$$M \gtrless \frac{\Delta_t^2}{2\sigma_t^2}. \tag{43}$$

The unique symmetric outcome is thus $\sigma_t = \sigma_L$ when $\Delta_t^2/2M < \sigma_L^2$, and $\sigma_t = \sigma_H$ when $\Delta_t^2/2M > \sigma_H^2$. When $\Delta_t^2/2M \in (\sigma_L^2, \sigma_H^2)$, on the other hand, firms mix between σ_L and σ_H, in proportions such that the resulting factor prices make each one indifferent; this equilibrium will be denoted σ_{LH}.[45] Focussing now on technology-inequality steady states, for any $\tau \leq 1$ and $\sigma \geq 1$ the *marginal benefit of flexibility* (right-hand-side of (43)) equals

$$R(\tau, \sigma) \equiv \frac{D(\tau; (\sigma - 1)/\sigma)^2}{2\sigma^2},$$

where $D(\tau, \gamma)$ is the asymptotic variance under the policy τ and return to skill γ, given by (21). Thus, under any time-invariant policy τ, whether exogenous or endogenous:

- For $M > \max\{R(\tau, \sigma_L), R(\tau, \sigma_H)\}$, the unique technological steady state is σ_L;
- For $M < \min\{R(\tau, \sigma_L), R(\tau, \sigma_H)\}$, it is σ_H;
- If $R(\tau, \sigma_L) > R(\tau, \sigma_H)$, then for $M \in [R(\tau, \sigma_H), R(\tau, \sigma_L)]$ it is the mixed-strategy outcome σ_{LH};
- If $R(\tau, \sigma_L) < R(\tau, \sigma_H)$, then for $M \in [R(\tau, \sigma_L), R(\tau, \sigma_H)]$ there are three technological steady states:

[44] The long-run multiplier for any shock to the $\mathcal{D}$ function (e.g., a change in w^2) is $\mu \equiv \left(1 - \mathcal{D}_1 - \mathcal{D}_2\left(\frac{\partial T}{\partial \Delta} + \frac{\partial T}{\partial \Gamma}\frac{\partial \Gamma}{\partial \Delta}\right) - \mathcal{D}_3\frac{\partial \Gamma}{\partial \Delta}\right)^{-1}$. Similarly, the long-run effects on inequality of a shock to the T function (e.g., a change in λ) its is $\mu \cdot \mathcal{D}_2(\partial T/\partial \lambda)$.

[45] It is not necessary to provide here the full characterization of this mixed-strategy equilibrium.

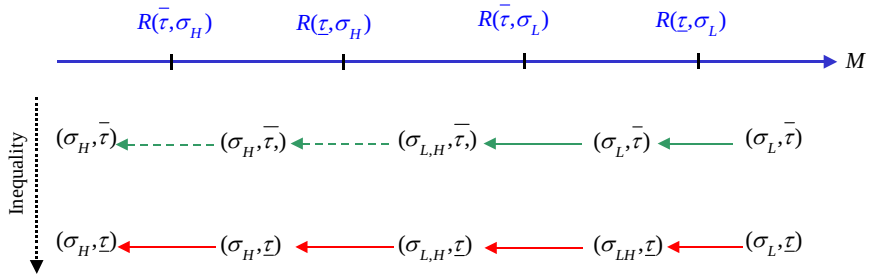


Figure 5: The response of technology and policy to a decline in the cost of flexibility (case (i)). Under each range of M appears the unique (σ,τ) such that $(\sigma,\Delta=D(\tau;1-1/\sigma))$ is a stable steady state given τ and M. The subset reached via solid lines corresponds to the stable steady-states in (σ,Δ,τ) jointly, when policy is endogenous as well.

σ_L, σ_H, and σ_{LH}; the first two are stable, the third one unstable.

Furthermore, since $R(\tau,\sigma)$ is decreasing in τ, we have:

Proposition 9 *More skill-biased technologies appear first in, and less skill biased technologies disappear first from, countries that have less redistributive fiscal, educational or labor market institutions. For any* $M>0$:

1) If σ_H *is a steady state equilibrium technology under a constant redistributive policy* τ, *this remains true under any less progressive policy* $\tau'<\tau$.

2) If σ_L *is a steady state equilibrium technology under a constant redistributive policy* τ', *this remains true under any more progressive policy* $\tau<\tau'$.

These results are illustrated in Figures 5 and 6 for two cases where: i) $R(\underline{\tau},\sigma_H)<R(\bar{\tau},\sigma_L)$, implying that for each M there is a unique technology compatible in the long-run with each policy $\tau\in\{\underline{\tau},\bar{\tau}\}$;[46] ii) $R(\underline{\tau},\sigma_L)<R(\bar{\tau},\sigma_H)$, implying that for either policy $\tau\in\{\underline{\tau},\bar{\tau}\}$ there is a range of M's where multiple technologies are sustainable. The message is essentially the same in both cases, showing how a *world-wide shift* in the set of feasible technologies can result in *different evolutions* of both production processes and the skill premium across countries. In particular, the model can help explain why skill-biased technical change and reorganization occurred first, and to a greater extent, in the United States compared to Europe –and within Europe, more so in England than on the Continent.[47]

[46] For instance, under condition (37), $\partial\ln\Delta_\infty/\partial\ln\sigma<1$, so $R(\underline{\tau},\sigma_H)<R(\underline{\tau},\sigma_L)$ provided σ_H and σ_L are close enough. If $\underline{\tau}$ and $\bar{\tau}$ are also not too different, then $R(\bar{\tau},\sigma)\precsim R(\underline{\tau},\sigma)$ for $\sigma=\sigma_H,\sigma_L$, so the thresholds rank as illustrated on Figure 5.

[47] Acemoglu (2003) proposes a different mechanism, based on imperfectly competitive labor markets, through which the wage-compression policies of continental European countries may have caused technological progress there to be less skill-biased than in the United States. In his model, a binding minimum wage makes low-skill workers' compensation a fixed price, whereas for high-skill workers the binding constraint for the firm is rent-sharing (due to search market frictions), which acts as a tax on productivity improvements. As a result, firms in high minimum-wage countries have greater incentives to invest in technologies that are complementary to low-skill labor than high-skill labor. In both Acemoglu's and the present model, the effects of policy on technology are indirect, operating through either the distribution of skills or equilibrium wages. In

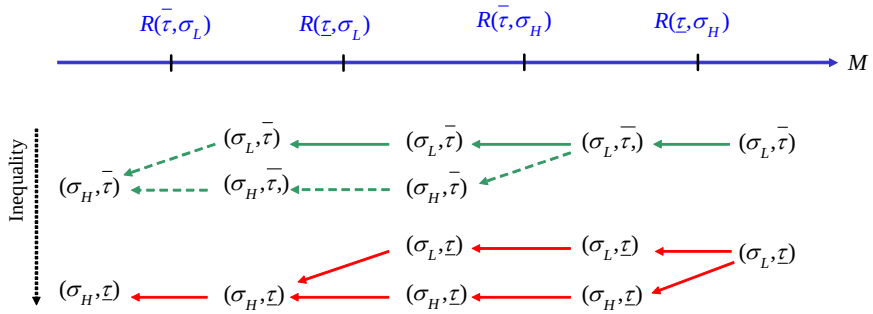


Figure 6: The response of technology and policy to a decline in the cost of flexibility (case (ii)). Under each range of M appear the values of (σ,τ) such that $(\sigma,\Delta=D(\tau;1-1/\sigma))$ is a stable steady state given τ and M. The subset reached via solid lines corresponds to the stable steady-states in (σ,Δ,τ) jointly, when policy is endogenous as well.

Indeed, consider two countries, C_1 and C_2, that are initially identical in all respects, including both using the technology σ_L, except that one is in a laissez-faire equilibrium, $\tau=\underline{\tau}$, and the other in a welfare state, $\tau=\bar{\tau}$. Suppose now that the technological frontier gradually flattens (M declines), meaning that flexibility becomes cheaper to achieve. As shown on Figures 5-6, the more skill-biased technology σ_H becomes (all or part of) another feasible equilibrium in C_1 before it does in C_2; similarly, σ_L first ceases to be viable (by itself or as part of a mixed equilibrium) in the laissez-faire country, while it is still sustainable in the more redistributive one.

Going further, there are in fact *reciprocal interactions* between the economy's *technology response* and *policy response* to shocks. Proposition 9 and Figures 5-6 show that feasible new technologies are not implemented unless institutions are sufficiently inegalitarian. But, conversely, the occurrence of technical change alters these same institutions, as seen in Proposition 7. Indeed, suppose that:

$$\underline{\lambda}(\gamma_L;B)<\lambda<\bar{\lambda}(\gamma_L;B), \tag{44}$$

where $\bar{\lambda}$ and $\underline{\lambda}$ were defined in (22)-(23). These inequalities imply that: i) under the technology σ_L, both social contracts $\underline{\tau}$ and $\bar{\tau}$ are political steady states; ii) under σ_H, $\bar{\tau}$ is a political steady state, while $\underline{\tau}$ is one if and only if we also have $\lambda<\bar{\lambda}(\gamma_H;B)$.

When this last inequality holds, the set of stable politico-economico-technological steady states (with endogenous τ,Δ and σ) is the same as described on Figures 5-6. When $\lambda>\bar{\lambda}(\gamma_H;B)$, however, the more redistributive social contract $\bar{\tau}$ is not *politically* sustainable under the amount of inequality that results, in the long run, from the technology σ_H. Therefore one must remove from the set of steady states on each figure the "branches" corresponding to this outcome; these are indicated by the dashed lines. The

Krusell and Rios-Rull (1996), by contrast, agents with different vintages of human capital vote directly on whether or not to allow the adoption of new technologies by firms.

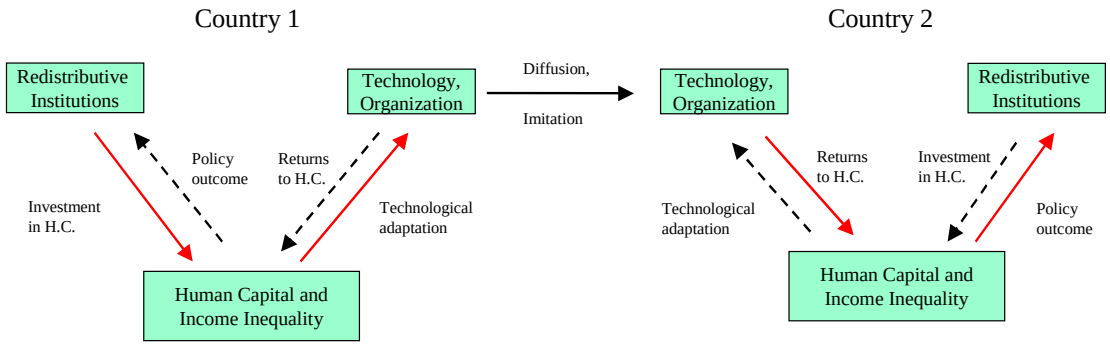


Figure 7: International spillovers between social contracts

remaining solid lines then indicate that *only certain politico-technological configurations* can be observed in the long run: a) for low values of M, e.g. for $M < R(\bar{\tau}, \sigma_L)$ on the first figure, the only feasible social contract is $\underline{\tau}$, together with the technology σ_H; b) on the second figure, for $M \in (R(\bar{\tau}, \sigma_L), R(\underline{\tau}, \sigma_L))$ only the egalitarian social contract and the egalitarian technology, or the inegalitarian social contract and inegalitarian technology, are mutually compatible.

V Exporting Inequality: Spillovers Between Social Contracts

The model also allows us to think about spillovers between national policies or institutions, operating via technological and organizational diffusion. The basic idea is illustrated in Figure 7, which shows how the social contract in Country 2 can, over time, be affected by technological or even purely political shifts in Country 1, propagated along the channels indicated by the solid lines on the diagram.

As seen in the previous section, firms operating in countries with more laissez-faire fiscal, educational or labor market policies have greater incentives to develop and adopt low-complementarity production processes. Suppose now that the cost of imitating, adapting or copying a more flexible technology or organizational form, once it has been developed and implemented elsewhere, is lower than the cost of innovation; in terms of the model, it is $m < M$. This lower marginal cost may for instance reflect, as in Acemoglu (1998), an imperfect international enforcement of property rights over technological or organizational innovations. As we shall see, redistributive institutions in one country will then be significantly affected, perhaps even completely undermined, by technological or political changes occurring in another.[48]

[48] As mentioned earlier I abstract here from international trade, which could be yet another channel of transmission. See Grossman and Maggi (2000), Grossman (2004) or Thoenig and Verdier (2003) for papers that study the effects of trade openness on technical and organizational change, although not their political economy implications.

A *A Shift in One Country's Technological Frontier*

I shall focus here on parameter configurations that satisfy the following conditions:

$$\max\left\{\underline{\lambda}(\gamma_L;B),\bar{\lambda}(\gamma_H;B)\right\} < \lambda < \bar{\lambda}(\gamma_L;B), \tag{45}$$

$$\max\left\{R(\underline{\tau},\sigma_L),R(\underline{\tau},\sigma_H)\right\} < M, \tag{46}$$

$$m < R(\bar{\tau},\sigma_L) < M' < \min\left\{R(\underline{\tau},\sigma_L),R(\underline{\tau},\sigma_H)\right\}, \tag{47}$$

which imply in particular that $M > M' > m$. As shown as part of Proposition 10 below, these conditions also ensure that the technology σ_L allows for both social contracts $\underline{\tau}$ and $\bar{\tau}$, and conversely that σ_L is an equilibrium technology under both social contracts (no firm wants to switch to σ_H).

Proposition 10 *Assume that conditions (45)–(47) are satisfied, and consider two countries, C_1 and C_2, that both start in steady state, with the same technology σ_L. Suppose now that the cost of flexibility in country C_1 declines from M to M'.*
1) If both C_1 and C_2 were initially in the more egalitarian of the two regimes compatible with σ_L nothing happens, in the sense that $(\bar{\tau},\gamma_L,D(\bar{\tau},\gamma_L))$ remains a stable steady state for both countries.
2) If C_1 was initially in the more inegalitarian regime $(\underline{\tau},\gamma_L,D(\underline{\tau},\gamma_L))$, the unique long run outcome is for both countries to switch to the technology σ_H, and for country C_2 to also adopt the more unequal social contract $\underline{\tau}$: the unique steady state for the two countries is now $(\underline{\tau},\gamma_H,D(\underline{\tau},\gamma_H))$.

The intuition is as follows. Even as M declines to M', firms faced with the skill distribution $D(\bar{\tau},\gamma_L)$ resulting from $\bar{\tau}$ do not find it profitable to switch technology. Given the higher dispersion $D(\underline{\tau},\gamma_L)$ that prevails under $\underline{\tau}$, however, if country C_1 starts in this regime all firms there will eventually switch to technology σ_H.[49] Next, given the lower cost of flexibility m to which firms in C_2 now have access through imitation, σ_L is no longer viable there even under $\bar{\tau}$. And, in turn, with the higher income inequality that results in the long run from technology σ_H, the only politically sustainable social contract is $\underline{\tau}$.

These results make clear how technological change (a shift in the frontier) has significant effects only when it is *mediated* through specific institutions –namely, which social contract C_1 had adopted; and, conversely, how under such conditions it will then affect institutions in other countries, namely here in C_2.

B *A Shift in One Country's Political Institutions*

I consider now a second scenario, namely the transmission of *a political shock.* Having seen earlier how the mere fact of being in different institutional steady states (say, for historical reasons) can lead to

[49] I leave aside the dynamics here, but they are straightforward: since (45) implies that (25) holds for $\gamma = \gamma_L$, γ_H, we always operate on the portion of the $T(\gamma\Delta)$ curve where increases in inequality imply decreases in the tax rate.

very different technological trajectories, I shall assume here that C_1 and C_2 both start in the egalitarian steady state, $(\bar{\tau}, \gamma_L, D(\bar{\tau}, \gamma_L))$, with the same technology σ_L. Let C_1 now experience an increase in the political influence of wealth, from λ to λ'. This may reflect a rising importance of *lobbying and campaign contributions*, an exogenous *decline in unionization,* or a lower *electoral turnout* by the poor. It may even simply represent the political outcome during a particular period in which the electorate stochastically shifted to the right.[50] I shall assume here the following conditions:

$$\bar{\lambda}(\gamma_H; B) < \lambda < \bar{\lambda}(\gamma_L; B) < \lambda', \tag{48}$$

$$m < R(\bar{\tau}, \sigma_L) < M < \min\left\{R(\underline{\tau}, \sigma_L), R(\underline{\tau}, \sigma_H)\right\}. \tag{49}$$

Proposition 11 *Assume that conditions (48)–(49) are satisfied. Consider two countries, C_1 and C_2, that both start in the egalitarian steady state, $(\bar{\tau}, \gamma_L, D(\bar{\tau}, \gamma_L))$, with the same technology σ_L. Suppose now that the political influence of wealth in country C_1 rises from λ to λ'. The unique long run outcome is for <u>both</u> countries to switch to the technology σ_H <u>and</u> the more unequal social contract $\underline{\tau}$, thus ending up at the steady state $(\underline{\tau}, \gamma_H, D(\underline{\tau}, \gamma_H))$.*

As a result of the initial political shift, redistribution τ_1 (fiscal, educational, or via labor-market institutions) in country C_1 declines. This leads over time to a rise in human capital inequality Δ_1, to which firms respond by adopting more flexible, wage-disequalizing technologies, switching from σ_L to σ_H and further precipitating the shift from $\bar{\tau}$ to $\underline{\tau}$. Their counterparts in C_2, which would not have developed such technologies by themselves, now find it profitable to copy them from C_1. This results in a rise in income inequality $\gamma_2\Delta_2$ in C_2 (and, over time, in human-capital inequality Δ_2 itself) that ultimately leads to the unravelling of the Welfare State in that country as well.

VI Conclusion

This chapter offers a new, unified model to analyze the reciprocal interactions between the distribution of human wealth, technology, and redistributive institutions. It identifies in particular certain core mechanisms that allow alternative societal models to persist, as well as powerful forces pushing towards uniformization. Key among the former is the interplay of imperfections in asset markets and in the political system that can lead to multiple steady states where inequality and redistribution are negatively correlated. Among the latter is skill-biased technical change, which can potentially lead to the unravelling of the Welfare State. When technological or organizational form is endogenous, moreover, firms respond to greater human capital heterogeneity with more flexible technologies, further exacerbating income inequal-

[50] Indeed, the political shock need not be permanent, provided the speed at which λ reverts to its previous value is low enough, relative to those of human capital adjustment and technological or organizational evolution.

ity. The possibility for firms in different countries to thus choose technologies adapted to the local labor force can also make it easier to sustain multiple social models. The international diffusion of technology, however, implies that the more flexible, skill-biased technologies profitably developed in nations with greater inequality and less redistributive institutions may then be imitated by firms in other countries, thereby triggering a "chain reaction" that moves the whole system towards a common outcome that is more inegalitarian –technologically, economically, and politically speaking. Such international spillovers between national social contracts are important concerns in the debate over globalization, and warrant further research.

Appendix

Proofs of Propositions 1-5. See Bénabou (2000); I shall only provide here:

(i) the formula for the break-even income level $\tilde{y}_t$ where $\hat{y}_t^i = y_t^i$,

$$\ln \tilde{y}_t = \gamma m_t + \delta \ln l_t + (2-\tau_t)\gamma^2 \Delta_t^2/2 \; + (1-\tau_t)v^2/2; \tag{A.1}$$

(ii) the laws of motion for (m_t, Δ_t^2) that underlie Proposition 2,

$$\begin{aligned} m_{t+1} &= (\alpha+\beta\gamma)\, m_t + \beta\delta \ln l_t + \beta\, \tau_t\, (2-\tau_t)\, (\gamma^2\Delta_t^2 + v^2)/2 + \ln\left(\kappa s^\beta\right) - (w^2+\beta v^2)/2 \quad \text{(A.2)} \\ \Delta_{t+1}^2 &= (\alpha+\beta\gamma(1-\tau_t))^2\, \Delta_t^2 + \beta^2(1-\tau_t)^2\, v^2 + w^2; \quad \text{(A.3)} \end{aligned}$$

(iii) and the formula for each agent's intertemporal welfare that underlies Proposition 3: under a rate of redistribution τ_t,

$$U_t^i = \bar{u}_t + A(\tau_t)(\ln k_t^i - m_t) + C(\tau_t) - (1-\rho+\rho\beta)(1-\tau_t)^2 \left(\gamma^2\Delta_t^2 + Bv^2\right)/2, \tag{A.4}$$

where $\bar{u}_t$ is independent of the policy τ_t, $B \equiv a + \rho(1-a)(1-\beta)$ was defined in (14) and:

$$\begin{aligned} A(\tau) &\equiv \rho\alpha + (1-\rho+\rho\beta)\gamma(1-\tau), \quad \text{(A.5)} \\ C(\tau) &\equiv (1-\rho)(\delta \ln l(\tau) - l(\tau)^\eta) + \rho\beta\delta \ln l(\tau), \quad \text{(A.6)} \end{aligned}$$

The first-order condition (15) readily follows. ■

Proof of Proposition 7. Because $D(\tau,\gamma)$ is increasing in γ for all τ the functions $\bar{\lambda}(\gamma;B)$ and $\underline{\lambda}(\gamma;B)$ are both U-shaped in γ, and minimized at the point where $\gamma D(\tau,\gamma) = v\sqrt{B}$, for $\tau = \bar{\tau},\ \underline{\tau}$ respectively. Furthermore, the minimum of $\bar{\lambda}(\gamma;B)$ occurs to the right of that of $\underline{\lambda}(\gamma;B)$. Under the assumption that $v\sqrt{B} > \gamma_{\max}\, D(\underline{\tau}, \gamma_{\max})$ we have $\gamma D(\bar{\tau},\gamma) < \gamma D(\underline{\tau},\gamma) < v\sqrt{B}$ for all $\gamma \leq \gamma_{\max}$, implying that both $\bar{\lambda}(\gamma;B)$ and $\underline{\lambda}(\gamma;B)$ are decreasing in γ over $[0, \gamma_{\max}]$; they are obviously increasing in B. Inverting these functions with respect to γ yields the claimed properties of $\underline{\gamma}(\lambda;B)$ and $\bar{\gamma}(\lambda;B)$. ■

Proof of Proposition 8. Consider a firm $\hat{\imath} \in [0,1]$ with technology $\hat{\sigma}$ and associated productivity factor $\hat{A} \equiv A(\hat{\sigma})$. Its marginal cost is:

$$MC(\hat{\sigma}|\sigma_t) \equiv \min_{\{\hat{x}_t^i\}} \left\{ \int_0^1 p_t^i\, \hat{x}_t^i\, di \;\middle|\; \hat{A} \cdot \left(\int_0^1 z_t^i \; \left(\hat{x}_t^i\right)^{\frac{\hat{\sigma}-1}{\hat{\sigma}}} \, ds \right)^{\frac{\hat{\sigma}}{\hat{\sigma}-1}} = 1 \right\}. \tag{A.7}$$

The first-order condition for cost-minimization is:

$$
\begin{aligned}
p_t^i &= \hat{\mu}_t\ \hat{A}\ z_t^i\ \left(\hat{x}_t^i\right)^{\frac{-1}{\hat{\sigma}}} \cdot \left(\int_0^1 z_t^i\ \left(\hat{x}_t^i\right)^{\frac{\hat{\sigma}-1}{\hat{\sigma}}}\ d\,s\right)^{\frac{1}{\hat{\sigma}-1}} = \hat{\mu}_t\ \hat{A}\ z_t^i\ \left(\hat{x}_t^i\right)^{\frac{-1}{\hat{\sigma}}} \cdot \left(\hat{A}\right)^{-\frac{1}{\hat{\sigma}}}, \quad \text{or:} \\
\hat{x}_t^i &= \hat{\mu}_t^{\hat{\sigma}}\ \hat{A}^{\hat{\sigma}-1} \left(\frac{p_t^i}{z_t^i}\right)^{-\hat{\sigma}}.
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
\hat{\mu}_t &= \int_0^1 p_t^i\, \hat{x}_t^i\, di = \hat{\mu}_t^{\hat{\sigma}}\ \hat{A}^{\hat{\sigma}-1} \left(\int_0^1 z_t^i\ \left(p_t^i/z_t^i\right)^{1-\hat{\sigma}}\, di\right), \quad \text{or:} \\
\hat{\mu}_t &= \hat{A}^{-1} \left(\int_0^1 \left(z_t^i\right)^{\hat{\sigma}} \left(p_t^i\right)^{1-\hat{\sigma}}\, di\right)^{\frac{1}{1-\hat{\sigma}}}, \qquad \text{(A.8)}
\end{aligned}
$$

which establishes (32). Now, replacing the equilibrium prices from equation (27) yields:

$$
\hat{\mu}_t = \hat{A}^{-1} A_t^{\frac{\sigma_t-1}{\sigma_t}} \left(\frac{y_t}{l_t}\right)^{\frac{1}{\sigma_t}} \left(\int_0^1 z_t^i\ \left(k_t^i\right)^{\frac{\hat{\sigma}-1}{\sigma_t}}\, di\right)^{\frac{1}{1-\hat{\sigma}}} = \hat{A}^{-1} A_t^{\frac{\sigma_t-1}{\sigma_t}} \left(\frac{y_t}{l_t}\right)^{\frac{1}{\sigma_t}} \left(\int_0^1 \left(k_t^i\right)^{\frac{\hat{\sigma}-1}{\sigma_t}}\, di\right)^{\frac{1}{1-\hat{\sigma}}},
$$

since the z_t^i's and k_t^i's are independent. We now eliminate the terms common to all firms by computing firm $\hat{\imath}$'s relative marginal cost:

$$
mc(\hat{\sigma}|\sigma_t) \equiv \frac{\hat{\mu}_t}{\mu_t} = \left(\frac{A_t}{\hat{A}}\right) \cdot \left(\int_0^1 \left(k_t^i\right)^{\frac{\hat{\sigma}-1}{\sigma_t}}\, di\right)^{\frac{1}{1-\hat{\sigma}}} \cdot \left(\int_0^1 \left(k_t^i\right)^{\frac{\sigma_t-1}{\sigma_t}}\, di\right)^{\frac{-1}{1-\sigma_t}}.
$$

Finally, using the fact that $\ln \int_0^1 \left(k_t^i\right)^{\chi} di = \exp\left[\chi m_t + \chi^2 \Delta_t^2/2\right]$ for all χ, this yields:

$$
mc(\hat{\sigma}|\sigma_t) = \left(\frac{A(\sigma_t)}{A(\hat{\sigma})}\right) \exp\left[\frac{\Delta_t^2}{2}\left(\frac{1-\hat{\sigma}}{\sigma_t^2} - \frac{1-\sigma_t}{\sigma_t^2}\right)\frac{\Delta_t^2}{2}\right] = \left(\frac{A(\sigma_t)}{A(\hat{\sigma})}\right) \exp\left[\frac{\Delta_t^2}{2}\left(\frac{\sigma_t-\hat{\sigma}}{\sigma_t^2}\right)\right]. \qquad \text{(A.9)}
$$

The (necessary and sufficient) first-order condition for firm $\hat{\imath}$ is therefore: $c'(\hat{\sigma}) = \Delta_t^2/2\sigma_t^2$. Evaluating it at $\hat{\sigma} = \sigma_t$ yields the technology-equilibrium condition $\sigma_t^2 c'(\sigma_t) = \Delta_t^2/2$, which by convexity of $c(\cdot)$ has a unique solution $\sigma^*(\Delta_t)$, increasing in Δ_t. Finally, the result that $\partial \ln \sigma^*/\partial \ln \Delta \in (0,1)$ is established in equation (35). ■

Proof of Section III.C's claims concerning growth with endogenous technology. In the general growth formula (12), $\delta \ln l_t$ is now replaced everywhere (according to (28)) by

$$
\ln \tilde{A}_t + \delta_t \ln l_t = \ln l_t + \ln\left(A_t^{\frac{\sigma_t-1}{\sigma_t}} (y_t)^{\frac{1}{\sigma_t}}\right) = \delta_t \ln l_t + \gamma_t \ln A(\sigma_t) + (1-\gamma_t) \ln y_t.
$$

This leads to:

$$\begin{aligned}\gamma_t \ln(y_{t+1}/y_t) &= \gamma_t \left[\ln\kappa + \beta \ln s + \delta_{t+1} \ln l_{t+1} - \alpha\delta_t \ln l_t + \ln A\left(\sigma_{t+1}\right) - \alpha \ln A\left(\sigma_t\right)\right] \\ &\quad -\gamma_t(1-\gamma_t)w^2/2 - \beta\gamma_t(1-\beta\gamma_t)(1-\tau_t)^2 v^2/2 \\ &\quad - \left[\alpha + \beta\gamma_t(1-\tau_t)^2 - (\alpha + \beta\gamma_t(1-\tau_t))^2\right]\gamma_t^2\Delta_t^2/2. \end{aligned} \tag{A.10}$$

For $\alpha = 0$ and $\beta = 1$, this simplifies to:

$$\begin{aligned}\gamma_t \ln(y_{t+1}/y_t) &= \gamma_t \left[\ln\kappa + \ln s + \delta_{t+1} \ln l_{t+1} + \ln A\left(\sigma_{t+1}\right)\right] \\ &\quad -\gamma_t(1-\gamma_t)\left[w^2/2 + (1-\tau_t)^2 v^2/2 + (1-\tau_t)^2\gamma_t^2\Delta_t^2/2\right] \\ &= \gamma_t \left[\ln\kappa + \ln s + \ln l_{t+1} + \ln A\left(\sigma_{t+1}\right)\right] \\ &\quad -\gamma_t(1-\gamma_t)\left[w^2/2 + (1-\tau_t)^2 v^2/2 + (1-\tau_t)^2\gamma_t^2\Delta_t^2/2\right], \end{aligned} \tag{A.11}$$

or, finally, since $\Delta_{t+1}^2 = \gamma^2(1-\tau_t)^2\,\Delta_t^2 + (1-\tau_t)^2\,v^2 + w^2$:

$$\begin{aligned}\ln(y_{t+1}/y_t) &= \ln\kappa + \ln s + \delta_{t+1}\ln l_{t+1} + \ln A\left(\sigma_{t+1}\right) - (1-\gamma_t)\Delta_{t+1}^2/2 \\ &= \ln\kappa + \ln s + \ln l_{t+1} - c(\sigma_{t+1}) - \frac{\Delta_{t+1}^2}{2\sigma_t}. \end{aligned} \tag{A.12}$$

Substituting for Δ_{t+1}^2 from (11), the growth rate between t and $t+1$ is thus:

$$g_t = \ln\kappa + \ln s + \delta_{t+1}\ln l(\tau_{t+1}) - c(\sigma_{t+1}) - \frac{\mathcal{D}(\Delta_t, \tau_t; \gamma_t)^2}{2\sigma_t}.$$

Therefore, with fixed labor supply ($1/\eta = 0$), if all firms are forced to use technology $\sigma_t - d\sigma$ instead of σ_t in every period the impact on growth will be $d\sigma$ times

$$\begin{aligned}\left.\frac{\partial g_t}{\partial\sigma}\right|_{\sigma=\sigma_t} &= -c'\left(\sigma_{t+1}\right) + \frac{\Delta_{t+1}^2}{2\sigma_t^2} - \frac{\Gamma'(\sigma_t)}{2\sigma_t}\cdot\frac{\partial\mathcal{D}^2(\Delta_t,\tau_t;\gamma_t)}{\partial\gamma} \\ &= -\frac{\Delta_{t+1}^2}{2}\left(\frac{\sigma_t^2}{\sigma_t^2} - 1\right) - \frac{\Gamma'(\sigma_t)}{2\sigma_t}\cdot\frac{\partial\mathcal{D}^2(\Delta_t,\tau_t;\gamma_t)}{\partial\gamma}, \end{aligned}$$

where we have used the condition for equilibrium technology choice in Proposition (8). The growth impact is thus positive in all periods provided that either $\sigma_{t+1}^2 \approx \sigma_t^2$ (we start in or near the steady state), or $\sigma_{t+1}^2 \leq \sigma_t^2$ (we start with "excessive" heterogeneity with respect to the steady state). ■

Proof of Proposition 10. We begin with some preliminaries. Given a technology σ and associated $\gamma = (\sigma-1)/\sigma$, recall from (22)-(23) that the tax rate $\bar{\tau}$ is a steady-state political equilibrium, which we shall denote as $\bar{\tau} \in \mathcal{P}(\sigma;\lambda)$, if and only if $\lambda \leq \bar{\lambda}(\gamma, B)$. Similarly, $\underline{\tau} \in \mathcal{P}(\sigma;\lambda)$ if and only if $\lambda \geq \underline{\lambda}(\gamma, B)$.

Conversely, given a tax rate τ we see from (43) that the technology σ_L and associated $\gamma_L = (\sigma_L - 1)/\sigma_L$ is a technological steady state when the slope of the technology frontier is M, which we denote as $\sigma_L \in T(\tau; M)$, if and only if:

$$M \geq \frac{D(\tau; \gamma_L)^2}{2\sigma_L^2} \equiv R(\tau, \sigma_L). \tag{A.13}$$

Conversely, the technology σ_H and associated $\gamma_H = (\sigma_H - 1)/\sigma_H$ is a technological steady state, which we denote as $\sigma_H \in T(\tau; M)$, if and only if:

$$M \leq \frac{D(\tau; \gamma_H)^2}{2\sigma_H^2} \equiv R(\tau, \sigma_H). \tag{A.14}$$

A policy-technology combination $(\tau, \sigma) \in \{\bar{\tau}, \underline{\tau}\} \times \{\sigma_L, \sigma_H\}$ is then a full steady-state if and only if $\tau \in P(\sigma; \lambda)$ and $\sigma \in T(\tau; M)$. Clearly, there are at most four stable steady-states (we restrict attention here to cases where the technology equilibrium is in pure strategies). We now proceed through a sequence of three claims, which together establish the proposition.

Claim 1: for a country facing the technological frontier M, the only steady states are $(\bar{\tau}, \sigma_L)$ and $(\underline{\tau}, \sigma_L)$. Indeed, the first inequality in (45) states that $\sigma_L \in T(\underline{\tau}; M)$ and this is easily seen to imply that $\sigma_L \in T(\bar{\tau}; M)$. Conversely, the second inequality states that $\sigma_H \notin T(\underline{\tau}; M)$ and this is easily seen to imply that $\sigma_L \in T(\bar{\tau}; M)$. Finally, the fact that $\underline{\lambda}(\gamma_L; B) < \lambda < \bar{\lambda}(\gamma_L; B)$ due to (45) means that $\underline{\tau} \in P(\sigma_L; \lambda)$ and $\bar{\tau} \in P(\sigma'_L; \lambda)$.

Claim 2: for a country facing the technological frontier M', the only steady states are $(\bar{\tau}, \sigma_L)$ and $(\underline{\tau}, \sigma_H)$. Indeed, note first from (47) that $R(\bar{\tau}, \sigma_L) < M'$ means that we still have $\sigma_L \in T(\bar{\tau}; M')$; by contrast, $M' < \min\{R(\underline{\tau}, \sigma_L), R(\underline{\tau}, \sigma_H)\}$ means that $\sigma_H \in T(\underline{\tau}; M)$ but $\sigma_L \notin T(\bar{\tau}; M)$. The only possible equilibria are thus $(\bar{\tau}, \sigma_L)$, $(\underline{\tau}, \sigma_H)$ and $(\underline{\tau}, \sigma_H)$. Turning now to (45), the fact that $\lambda < \bar{\lambda}(\gamma_L; B)$ means that $\bar{\tau} \in P(\sigma'_L; \lambda)$; the fact that that $\bar{\lambda}(\gamma_H; B) < \lambda$, on the other hand, means that $\underline{\tau} \in P(\sigma_H; \lambda)$ but $\underline{\tau} \notin P(\sigma_H; \lambda)$. So only the first two of the three preceding configurations are full equilibria.

Claim 3: *for a country facing the technological frontier m, the only steady state is $(\underline{\tau}, \sigma_H)$.* Observe from (47) that m satisfies all the same inequalities as M', except that $m < R(\bar{\tau}, \sigma_L)$ whereas $R(\bar{\tau}, \sigma_L) < M'$. This means that whereas we had $\sigma_L \in T(\bar{\tau}; M')$, we now have $\sigma_L \notin T(\bar{\tau}; M')$. This rules out the equilibrium $(\bar{\tau}, \sigma_L)$, leaving only $(\underline{\tau}, \sigma_H)$. ■

Proof of Proposition 11.

Claim 1: in the initial parameter configuration, $(\bar{\tau}\ , \sigma_L)$ is a steady state (and even the only steady-state with policy $\bar{\tau}$). Indeed, the fact that $\bar{\lambda}(\gamma_H; B) < \lambda < \bar{\lambda}(\gamma_L; B)$ means that $\bar{\tau} \in \mathcal{P}(\sigma_L; \lambda)$, whereas $\bar{\tau} \notin \mathcal{P}(\sigma_H; \lambda)$. The rest of the claim follows from the fact $\sigma_L \in \mathcal{T}(\bar{\tau}; M)$, since $M > R(\bar{\tau}, \sigma_L)$.

Claim 2: After the political shift in C_1, $(\underline{\tau}, \sigma_H)$ is the only steady-state for that country. First, since $\lambda' > \bar{\lambda}(\gamma_L; B) > \bar{\lambda}(\gamma_H; B)$ we now have $\bar{\tau} \notin \mathcal{P}(\sigma_L; \lambda')$ and $\bar{\tau} \notin \mathcal{P}(\sigma_H; \lambda')$, so there is no steady-state with policy $\bar{\tau}$. Moreover, since $M < R(\underline{\tau}, \sigma_L)$ we have $\sigma_L \notin \mathcal{T}(\underline{\tau}; M)$, so the only possible equilibrium is $(\underline{\tau}, \sigma_H)$. It is indeed an equilibrium, as $M < R(\underline{\tau}, \sigma_H)$ means that $\sigma_H \in \mathcal{T}(\underline{\tau}; M)$, while $\bar{\lambda}(\gamma_H; B) < \lambda'$ means that $\underline{\tau} \in \mathcal{P}(\sigma_H; \lambda')$.

Claim 3: After C_1 has switched to the technology σ_H, so that C_2 faces the technology frontier m, the only steady-state for C_2 is $(\underline{\tau}, \sigma_H)$. First the fact $m < R(\bar{\tau}, \sigma_L) < R(\underline{\tau}, \sigma_L)$ implies that $(\bar{\tau}, \sigma_L)$ is no longer a technological equilibrium, and a fortiori neither is $(\underline{\tau}, \sigma_L)$. Second, the fact $m < \min\{R(\underline{\tau}, \sigma_L), R(\underline{\tau}, \sigma_H)\}$ means that the only technological equilibrium under policy $\underline{\tau}$ is σ_H. Finally, since $\lambda > \bar{\lambda}(\gamma_H; B)$, $\underline{\tau} \in \mathcal{P}(\sigma_H; \lambda)$ whereas $\bar{\tau} \notin \mathcal{P}(\sigma_H; \lambda)$, which concludes the proof. ■

References

Acemoglu, D. (1998) "Why Do New Technologies Complement Skills? Directed Technical Change and Inequality," *Quarterly Journal of Economics,* 113(4), 1055-1090.

Acemoglu, D. (2003) "Cross-Country Inequality Trends," *Economic Journal,* 113, 121-149.

Acemoglu, D., Aghion P. and Violante G.L. (2001) "Deunionization, Technological Change and Inequality," *Carnegie-Rochester Conference Series on Public Policy,* 55, 229-264.

Aghion, P. (2002) "Schumpeterian Growth Theory and the Dynamics of Income Inequality," *Econometrica,* 70(3), 855-882.

Alesina, A., Glaeser, E. and Sacerdote, B. (2002) "Why Doesn't the US Have a European-Type Welfare State?" *Brookings Papers on Economic Affairs,* Issue 2, 187-277.

Alesina, A. and Rodrik, D. (1994) "Distributive Politics and Economic Growth." *Quarterly Journal of Economics,* 109(2), 465-490.

Alesina, A. and Angeletos, G.M. (2003) "Fairness and Redistribution: US vs. Europe," NBER Working Paper 9502, February.

Autor, D., L. Katz and A. Krueger (1997) "Computing Inequality: Have Computers Changed the Labor Market?", *Quarterly Journal of Economics,* 113(4), 1169-1214.

Baland, J.M. and Robinson, J. (2003) "Land and Power," University of California Berkeley, mimeo.

Banerjee, A. and Duflo, E. (2000) "Inequality and Growth: What Can the Data Say?", *Journal of Economic Growth,* 8(3), 267-299.

Banerjee, A. and Newman**,** A. (1993) "Occupational Choice and the Process of Development," *Journal of Political Economy,* 274-278.

Barro R. (2000) "Inequality and Growth in a Panel of Countries," *Journal of Economic Growth,* 5, 5-32.

Bartels, L. (2002) "Economic Inequality and Political Representation," Princeton University mimeo.

Becker (1964) *Human Capital.* Columbia University Press, New York.

Bénabou, R. (1996a) "Heterogeneity, Stratification, and Growth: Macroeconomic Implications of Community Structure and School Finance." *American Economic Review,* 86(3), 584–609.

Bénabou, R. (1996b) "Inequality and Growth," in Ben S. Bernanke and Julio J. Rotemberg eds., *National Bureau of Economic Research Macro Annual,* vol. 11, 11-74. Cambridge, MA: MIT Press.

Bénabou, R. (2000) "Unequal Societies: Income Distribution and the Social Contract," *American Economic Review,* 90, 96-129.

Bénabou, R. (2002) "Tax and Education Policy in a Heterogenous Agent Economy: What Levels of Redistribution Maximize Growth and Efficiency?", *Econometrica,* 70, 481-517.

Bénabou, R. and Tirole, J. (2002) "Belief in a Just World and Redistributive Politics". Institute for Advanced Study mimeo, October.

Berman, E., J. Bound and S. Machin (1997) "Implications of Skill-Biased Technological Change: International Evidence", *Quarterly Journal of Economics,* 113(4), 1245-1280.

Bisin, A. and Verdier, T. (1999) "Work Ethic and Redistribution: A Cultural Transmission Model of the Welfare State," DELTA mimeo, December.

Bourguignon, F. and Verdier, T. (2000) "Oligarchy, Democracy, Inequality and Growth," *Journal of Development Economics,* 62(2), 285-314.

Caroli, E. and Van Reenen, J. (1999) "Skill Biased Organizational Change? Evidence from a Panel of British and French Establishments," CEPREMAP Working Paper No. 9917.

De Mello, L. and Tiongson, E. (2003) "Income Inequality and Redistributive Government Spending," IMF Working Paper 314.

Engerman, S. and Sokoloff, K. (1997) "Factor Endowments, Institutions, and Differential Growth Paths among New World Economies," in S. Haber ed. *"How Latin America Fell Behind: Essays on the Economic Histories of Brazil and Mexico, 1800-1914,"* Stanford University Press, 260-304.

Fellman, J. (1976) "The Effect of Transformations on Lorenz Curves," *Econometrica,* 44 (4), 823-24.

Fernandez, R. and Rogerson, R. (1998) "Public Education and the Dynamics of Income Distribution: A Quantitative Evaluation of Education Finance Reform." *American Economic Review*, 88(4), 813–833.

Figini, P. (1999) "Inequality and Growth Revisited," Trinity College Economic Papers, Dublin, No. 2/99, September.

Forbes, Kristen. (2000) "A Reassessment of the Relationship Between Inequality and Growth", *American Economic Review,* 90(4), 869-887.

Fortin, N. and Lemieux, T. (1997) "Institutional Changes and Rising Wage Inequality: Is There A Linkage?, *Journal of Economic Perspectives,* 11(2), Spring, 75-97.

Freeman, R. (1995) "The Large Welfare State as a System," *American Economic Review Papers and Proceedings,* 85(2), 16-21.

Freeman, R. (1996) "Labor Market Institutions and Earnings Inequality," *New England Economic Review,* May/June, 169-172.

Galor, O. and Tsiddon, D. (1997) "Technological Progress, Mobility, and Growth," *American Economic Review,* 87 (June), 363-382.

Galor, O. and Moav, O. (2000) "Ability Biased Technological Transition, Wage Inequality, and Economic Growth," *Quarterly Journal of Economics* 113(4), 1055-1090.

Galor, O. and Moav, O. (1999) "From Physical to Human Capital: Inequality in the Process of Development," CEPR Discussion Paper No. 2307, December.

Galor, O. and Zeira, J. (1993) "Income Distribution and Macroeconomics." *Review of Economic Studies,* 60(1), 35-52.

Glomm, G. and Ravikumar, B. (1992) "Public vs. Private Investment in Human Capital: Endogenous Growth and Income Inequality," *Journal of Political Economy,* 100, 818-834.

Gradstein, M. and Justman, M. (1997) "Democratic Choice of an Education System: Implications for Growth and Income Distribution," *Journal of Economic Growth,* 2(2), 169–184.

Grossman, G. (2004) "The Distribution of Talent and the Pattern and Consequences of International Trade," *Journal of Political Economy,* 112(1), 209-239.

Grossman, G. and Maggi, G. (2000). "Diversity and Trade," *American Economic Review,* 90(5), 1255-1275.

Hassler J., and Rodriguez–Mora, J. (1999) "Employment Turnover and the Public Allocation of Unemployment Insurance," *Journal of Public Economics,* 73(1), 55-83.

Hassler J., Rodriguez–Mora, J., Storesletten, K., and Zillibotti, F. (2003) **"**The Survival of the Welfare State," *American Economic Review,* 93(1), 87-112.

Kiley, M. (1999) "The Supply of Skilled Labor and Skill-Biased Technological Progress," *Economic Journal,* 109, October, 08-724.

Kremer, M., and Maskin, E. (1996) "Wage Inequality and Segregation by Skill," NBER Working Paper No. 5718.

Kremer, M., and Maskin, E. (2003) "Globalization and Inequality," Harvard University mimeo.

Kreps, D. and Porteus, E. (1979) "Dynamic Choice Theory and Dynamic Programming," *Econometrica,* 47(1), 91–100.

Krueger, A. (2002) "Inequality, Too Much of a Good Thing?", Princeton University mimeo, April.

Krusell, P. and Rios-Rull, J.V. (1996) "Vested Interests In a Positive Theory of Stagnation and Growth," *Review of Economic Studies,* 63, 301-329.

Lee, D. (1999) "Wage Inequality in the United States During the 1980s: Rising Dispersion or Falling Minimum Wage?", *Quarterly Journal of Economics,* 114(3), 977–1023.

Lee, W. and Roemer, J. (1998) "Income Distribution, Redistributive Politics, and Economic Growth," *Journal of Economic Growth,* 3, 217-240.

Legros, P. and Newman A. (1996) "Wealth Effects, Distribution, and the Theory of Organization," *Journal of Economic Theory,* 70(2), 312-341.

Lindbeck, A. (1995) "Welfare-State Disincentives with Endogenous Habits and Norms," *Scandinavian Journal of Economics,* 97(4), 477-494.

Lloyd-Ellis, H. (1999) "Endogenous Technological Change and Inequality," *American Economic Review,* 89(1), 47-77.

Loury, G. (1981) "Intergenerational Transfers and the Distribution of Earnings," *Econometrica,* 49, 843-867.

Moav, O. (2002) "Income Distribution and Macroeconomics: the Persistence of Inequality in a Convex Technology Framework," *Economics Letters,* 75(2), 147-287.

Mookherjee, D. and Ray, D. (2003) "Persistent Inequality," *Review of Economic Studies,* 70, 369–393.

Perotti, R. (1996) "Growth, Income Distribution and Democracy: What the Data Say," *Journal of Economic Growth,* 1(2), 149–187.

Persson, T. and Tabellini, G. (1991) "Is Inequality Harmful for Growth? Theory and Evidence," *American Economic Review,* 84(3), 600-621.

Pineda, J. and Rodriguez, F. (2000) "The Political Economy of Human Capital Accumulation," University of Maryland at College Park mimeo, February.

Piketty, T. (1995) "Social Mobility and Redistributive Politics," *Quarterly Journal of Economics,* 110(3), 551–442.

Piketty, T. (1997) "The Dynamics of the Wealth Distribution and Interest Rate with Credit-Rationing," *Review of Economic Studies,* 64(2), 173–190.

Rodriguez, F. (1999) "Does Inequality Lead to Redistribution? Evidence from the United States," *Economics & Politics,* 11(2), 171-199.

Romer, P. (1987) "Growth Based on Increasing Returns Due to Specialization," *American Economic Review* 77, May, 56-62.

Rosenstone, S. and Hansen, J. (1993) *Mobilization, Participation and Democracy in America.* New York: MacMillan Publishing Company.

Saint-Paul, G. (2001) "The Dynamics of Exclusion and Fiscal Conservatism," Review of Economic Dynamics, 4, 275-302.

Saint-Paul, G., and Verdier, T. (1993) "Education, Democracy and Growth," *Journal of Development Economics,* 42(2), 399-407.

Sheshadri, A. and K. Yuki (2004) "Equity and Efficiency Effects of Redistributive Policies," *Journal of Monetary Economics,* 57(1), 1415-1447

Tamura, R. (1992) "Efficient Equilibrium Convergence: Heterogeneity and Growth," *Journal of Economic Theory,* 58(2), 55-76.

Thesmar, D. and Thoenig, M. (2000) "Creative Destruction and Firm Organizational Choice," *Quarterly Journal of Economics,* 115(4), 1201-1239.

Thoenig, M. and Verdier, T. (2003) **"**A Theory of Defensive Skill-Biased Innovation and Globalization," *American Economic Review,* 93(3), 709-728

Vindigni, A. (2002a) "Income Distribution and Skilled-Biased Technical Change," Université de Toulouse mimeo, June.

Zhang, J. (2004) "Income Ranking, Convergence Speeds, and Growth Effects of Inequality with Two-Dimensional Adjustment," *Journal of Economic Dynamics and Control,* forthcoming.

Social Capital

Steven N. Durlauf and Marcel Fafchamps*

April 28, 2004

*We thank Christian Grootaert for initiating this work and for helpful suggestions. Durlauf thanks the University of Wisconsin and John D. and Catherine T. MacArthur Foundation and Durlauf and Fafchamps thank the World Bank for financial support. Ritesh Banerjee, Ethan Cohen-Cole, Artur Minkin, Giacomo Rondina and Chih Ming Tan have provided excellent research assistance. Jim Magdanz has provided useful comments on an earlier draft.

JEL Classification Codes: E26, O10, O40, L14, Z13
Keywords: development, growth, identification, inequality, networks, social capital, trust

Social Capital

Abstract

This paper surveys research on social capital. We explore the concepts that motivate the social capital literature, efforts to formally model social capital using economic theory, the econometrics of social capital, and empirical studies of the role of social capital in various socioeconomic outcomes. While our focus is primarily on the place of social capital in economics, we do consider its broader social science context. We argue that while the social capital literature has produced many insights, a number of conceptual and statistical problems exist with the current use of social capital by social scientists. We propose some ways to strengthen the social capital literature.

Steven N. Durlauf
Department of Economics
University of Wisconsin
1180 Observatory Drive
Madison, WI 53706-1393
United States
sdurlauf@ssc.wisc.edu

Marcel Fafchamps
Department of Economics
University of Oxford
Manor Road
Oxford, OX1 3UQ
United Kingdom
marcel.fafchamps@economics.oxford.ac.uk

> ...in every community there seems to be some sort of justice, and some type of friendship, also. At any rate, fellow-voyagers and fellow-soldiers are called friends, and so are members of other communities. And the extent of their community is the extent of their friendship, since it is also the extent of the justice found there...What is just...is not the same for parents towards children as for one brother towards another, and the same for companions as for fellow-citizens, similarly with the other types of friendship...what is unjust towards of these is also different, and become more unjust as it is practiced on closer friends. It is more shocking, e.g., to rob a companion of money than to rob a fellow-citizen, to fail to help a brother than a stranger, and to strike one's father than anyone else. What is just also naturally increases with friendship, since it involves the same people and extends over an equal area.
>
> Aristotle, *Nicomachean Ethics*, Book VIII, 9.61

I. Introduction

Social capital represents one of the most powerful and popular metaphors in current social science research. Broadly understood as referring to the community relations that affect personal interactions, social capital has been used to explain an immense range of phenomena, ranging from voting patterns to health to the economic success of countries. Literally hundreds of papers have appeared throughout the social science literature arguing that social capital matters in understanding individual and group differences and further that successful public policy design needs to account for the effects of policy on social capital formation.

This paper is designed to survey research on social capital. We will give primary focus to the role of social capital in economic growth and development as suggested by the presence of this paper in the Handbook of Economic Growth. That being said, this survey will discuss social capital in general as there is no part of the social capital literature that may plausibly be treated as orthogonal to the issues that arise in relating social capital to economic growth. Our objectives are threefold. First, we provide an overview of conceptual issues that underlie social capital studies. Second, we identify

some general flaws we see in the empirical social capital literature. While we would hardly claim that every social capital study suffers from these problems, we do claim that they are prevalent in the literature. Third, we make a number of recommendations on how to strengthen the social capital literature. In assessing empirical work, we will focus almost exclusively on statistical analyses of social capital. This is not because we regard qualitative studies as unimportant (we will in fact advocate their greater use in the course of our discussion) but because such studies raise very distinct conceptual and interpretative questions from their quantitative counterparts.

Much of our discussion is critical. We argue that empirical social capital studies are often flawed and make claims that are in excess of what is justified by the statistical exercises reported. However, this should not be taken as an indictment of research on social capital per se. In our judgment the role of social factors in individual and group outcomes is of fundamental importance in most of the contexts in which social capital has been studied. Hence we regard the empirical social capital literature as addressing major outstanding issues in many areas of social science. Our intent in this survey is to evaluate what is currently known and to make suggestions on how to improve future research.

The paper is organized as follows. Section II contains a discussion of how economists and other social scientists have attempted to define social capital. The section also reviews some of the contexts in which social capital has been argued to play an important causal role in various sociological outcomes. Section III discusses efforts to theorize about social capital; both heuristic and conceptual arguments are discussed as well as formal analyses. Section IV discusses econometric issues that arise in the efforts to develop empirical evidence of the role of social capital as a determinant of socioeconomic outcomes. Section V reviews the empirical literature on social capital; while this literature is far too large to cover comprehensively we believe our survey captures the range of contexts in which social capital effects have been evaluated. Section VI reviews empirical studies that analyze the determinants of social capital. Section VII contains some suggestions for improving social capital research. Section VIII concludes.

II. Social capital: basic concepts

II.i. Defining social capital

Since Loury (1977) introduced it into modern social science research and Coleman's (1988) seminal study placed it at the forefront of research in sociology, the term social capital has spread throughout the social sciences and has spawned a huge literature that runs across disciplines. Despite the immense amount of research on it, however, the definition of social capital has remained elusive. From a historical perspective, one could argue that social capital is not a concept but a *praxis*, a code word used to federate disparate but interrelated research interests and to facilitate the cross-fertilization of ideas across disciplinary boundaries. The success of social capital as a federating concept may result from the fact that no social science has managed to impose a definition of the term that captures what different researchers mean by it within a discipline, let alone across fields.[1]

While conceptual vagueness may have promoted the use of the term among the social sciences, it also has been an impediment to both theoretical and empirical research of phenomena in which social capital may play a role.[2] In order to anchor our discussion of social capital, we need a substantive definition. We begin our search by listing a number of definitions that have been proposed by some of the most influential researchers on social capital. We begin with Coleman (1990) who defines social capital as:

[1]Even if a precise definition of social capital were attempted, it is likely to be no less vague than other similar concepts. The term capital, for instance, is used to describe different things – from finance to machinery to infrastructure. Human capital similarly has many different meanings, such as education, nutrition, health, vocational skills, and knowledge. This kind of vagueness, however, is less problematic as long as researchers agree on some basic principles.

[2] Criticisms of the vagueness and inconsistency of various definitions of social capital may be found in Dasgupta (2000), Durlauf (2000), Manski (2000) and Portes (1998). Arrow (2000) goes so far as to suggest that the term social capital be abandoned.

> ...social organization constitutes social capital, facilitating the achievement of goals that could not be achieved in its absence or could be achieved only at a higher cost. (pg. 304)

Putnam *et al* (1993) provides a similar characterization,

> ...social capital...refers to features of social organization, such as trust, norms, and networks that can improve the efficiency of society... (pg. 167)

Both definitions emphasize the beneficial effects social capital is assumed to have on social aggregates. According to these definitions, social capital is a type of positive group externality. Coleman's definition suggests that the externality arises from social organization. Putnam's definition emphasizes specific *informal* forms of social organization such as trust, norms and networks. In his definition of social capital, Fukuyama (1997) argues that only certain shared norms and values should be regarded as social capital:

> Social capital can be defined simply as the existence of a certain set of informal rules or norms shared among members of a group that permits cooperation among them. The sharing of values and norms does not in itself produce social capital, because the values may be the wrong ones... The norms that produce social capital... must substantively include virtues like truth-telling, the meeting of obligations, and reciprocity. (pp. 378-379)

Other definitions characterize social capital not in terms of outcome but in terms of relations or interdependence between individuals. In later research, Putnam (2000) defines social capital as

> ...connections among individuals - social networks and the norms of reciprocity and trustworthiness that arise from them. (pg. 19)

Ostrom (2000) writes

> *Social capital* is the shared knowledge, understandings, norms, rules and expectations about patterns of interactions that groups of individuals bring to a recurrent activity. (pg. 176)

In a similar vein Bowles and Gintis (2002) state

> Social capital generally refers to trust, concern for one's associates, a willingness to live by the norms of one's community and to punish those who do not. (pg. 2)

Finally, one finds in a recent book-length treatment, Lin (2001)

> ...social capital may be defined operationally as *resources embedded in social networks and accessed and used by actors for actions.* Thus, the concept has two important components: (1) it represents resources embedded in social relations rather than individuals, and (2) access and use of such resources reside with actors. (pp. 24-25)

From these definitions, we can distinguish three main underlying ideas: (1) social capital generates positive externalities for members of a group; (2) these externalities are achieved through shared trust, norms, and values and their consequent effects on expectations and behavior; (3) shared trust, norms, and values arise from informal forms of organizations based on social networks and associations. The study of social capital is that of network-based processes that generate beneficial outcomes through norms and trust.

By this definition social capital is always desirable since its presence is equated with beneficial consequences. This formulation is quite unsatisfactory from the perspective of policy evaluation (e.g., Durlauf (1999,2002b), Portes (1998)): if one denies the appellation of social capital to contexts where strong social ties lead to immoral or unproductive behaviors, there is nothing nontrivial to say in terms of policy. Presumably it is social structures, not their consequences, which can be influenced by policymakers.

Unless we know under what conditions social structures generate beneficial outcomes, we cannot orient policy. We also note that the benefits that social capital generates for one group may disadvantage another, so that the combined effect on society need not be positive. We come back to this issue later.

The three main ideas outlined above often appear intertwined in the mind of their proponents so that one in isolation would probably not be considered social capital. For instance, there are many phenomena that generate positive (or negative) externalities. According to the definitions listed here, they would probably not be considered social capital unless they involve norms or trust. There appears to be more confusion as to whether all three parts of the definition are required for social capital. Norms and trust can be based on formal institutions such as laws and courts without reference to social networks. Yet the literature sometimes has referred to such generalized trust as social capital (e.g., Knack and Keefer (1997)). It is also unclear whether (1) and (3) alone constitute social capital. In his seminal work on job markets, for instance, Granovetter (1975) discusses how social networks are activated to share job market information, thereby speeding job search and raising the efficiency of the job matching process. This process does not, by itself, require shared norms or values. Fafchamps and Minten (2002) use the phrase 'social network capital' to describe this phenomenon.

From the perspective of empirical work, a definition of social capital limited to (1) and (2) is problematic. Things like 'norms' and 'shared values' are notoriously difficult to measure. This has led some of the less rigorous work in this area to present evidence of a beneficial group effect as evidence of social capital itself, and consequently to conclude that social capital is good. This kind of circular reasoning is of course not satisfactory since it is ultimately tautological and is not falsifiable.

A definition of social capital suitable for rigorous empirical work must identify observable variables that can be used as proxies for social capital (Portes (2000)). Norms, trust, and expectations of behavior are very broad ideas that encompass no end of phenomena. Identifying a commonly acceptable set of proxies for social capital has therefore proved a formidable task and many different variables have appeared in empirical papers purportedly to measure it. Another problem has to do with the extent to which the variables used identify well defined social influences – part (3) of our

definition. Adherence to norms can be induced for many reasons, including many that cannot be reasonably construed as social. Consequently, evidence of adherence to norms does not, by itself, constitute evidence of the importance of social networks. To the extent that social networks and associations are part of the definition of social capital, evidence must also be provided that trust and shared norms are achieved via social interaction based on interpersonal networks and associations.

II.ii The efficiency of social exchange

Perhaps a more fruitful approach for our purpose is to proceed by example, that is, to select one specific phenomenon and use it to illustrate how research on social capital can be organized. Much of the commonality in definitions of social capital and in examples given by respective authors is the focus on interpersonal relationships and social networks and their effect on the efficiency of social exchange – whether the provision of a public good, as in Coleman's work, or the better organization of markets, as in Granovetter's. At the heart of the concept of social capital is the idea that positive externalities cannot be achieved without some kind of coordination, i.e., there is coordination failure. Much of the interest in social capital stems from efforts to understand how socially efficient outcomes can occur in environments in which the sorts of conditions necessary for the classical First Welfare Theorem are not fulfilled. Efficiency of social exchange is thus a good vantage point around which to organize our assessment.

One important potential role for social capital concerns its ability to ameliorate potential inefficiencies caused by imperfect information. As Hayek (1945) was among the first to point out, information asymmetries are an inescapable feature of human society. As a result, exchange is hindered either because agents who could benefit from trade cannot find each other, or because, having found each other, they do not trust each other enough to trade. In either case, some mutually beneficial exchange does not take place. Similar principles apply to the provision of public goods. Search and trust are thus two fundamental determinants of the efficiency of social exchange. If we can finds ways of facilitating search and of fostering trust, we can improve social exchange.

There are basically two ways of achieving these dual objectives: via formal institutions (e.g., a stock exchange or a trading fair) or via interpersonal relationships (e.g., word-of-mouth communication of opportunities, repeated interactions which benefit both parties). The literature on social capital focuses principally on the latter. In the following discussion, we illustrate how social networks can raise efficiency. We begin by examining the possible effects of social networks on search. In so doing, we focus only on parts (1) and (3) of our definition of social capital since norms and trust are not central to the circulation of information (although they can play a subsidiary role). We then turn to trust, the externalities it generates, and the way to sustain trust through social networks. Public goods are discussed in the following sub-section. The relationship between social capital and economic development is examined next. The last sub-section explores the relationship between social capital and equity.

Social networks and search

The role of social capital in search can be illustrated by comparing US equity and labor markets. Given the existence of a stock market, it is very easy for a seller of stock to find a buyer at the market clearing price. This is not the case in labor markets where no equivalent institution circulates accurate and up-to-date information about jobs and workers. In his path-breaking study of the US labor market, Granovetter (1975) brought to light the role played by interpersonal relationships in channeling information about jobs and job applicants. A large proportion of jobs are allocated on the basis of personal recommendation and word-of-mouth. This can be understood as an endogenous, spontaneous adaptation to the absence of a formal clearing house equivalent to the stock market.[3]

As this comparison demonstrates, observing that social capital plays a role in markets does not, by itself, constitute evidence that social capital is necessary and should

[3] This is not to say that efforts have not been made to emulate the stock market model – from employment offices to internet sites to temporary employment agencies. But to date none of these institutions seems capable of conveying sufficiently precise information about jobs and job applicants, especially regarding worker environment, work ethics, and

be nurtured. Depending on the circumstances, the development of formal institutions may be a superior alternative.

Social capital and trust

As argued in Fafchamps (2004), trust may be understood as an optimistic expectation or belief regarding other agents' behavior. The origin of trust may vary.[4] Sometimes, trust arises from repeated interpersonal interaction. Other times, it arises from general knowledge about the population of agents, the incentives they face, and the upbringing they have received (Platteau (1994a,b)). The former can be called personalized trust and the latter generalized trust. The main difference between the two is that, for each pair of newly matched agents, the former takes time and effort to establish while the latter is instantaneous.

In most situations, trusting others enables economic agents to operate more efficiently – e.g. by invoicing for goods they have delivered or by agreeing to stop hostilities. Whenever this is the case, generalized trust yields more efficient outcomes than personalized trust. The reason is that, for any pair of agents, generalized trust is established faster and more cheaply than personal trust. This observation has long been made in the anthropological literature on generalized morality. Fostering generalized trust can thus potentially generate large efficiency gains. How this can be accomplished, however, is unclear.

Clubs and networks are different concepts having to do with the structure of links among economic agents. Clubs describe finite, closed groupings. Networks describe more complex situations in which individual agents are related only to some other agents, not all. The term 'network' is sometimes used to describe the entire set of links among a finite collection of agents. Other times, it is used to describe the set of links around a specific individual. To avoid confusion, we refer to the second concept as a subjective network.

---

personal motivation. See Fafchamps (2002) and Kranton (1996) for models of spontaneous market emergence organized around interpersonal relationships.

[4] Sometimes trust is misplaced, but for the sake of brevity, we ignore this possibility here. Put differently, we assume rational expectations.

Among other things, clubs and networks can be used to describe the extent to which personalized and generalized trust exist in a population. Perfect generalized trust corresponds to the case where all agents belong to a single club (or complete network) and trust all other members. Situations in which generalized trust exists only among sub-populations (say, Jewish diamond dealers in New York, cf. Bernstein (1992)) could be described as small clubs. Situations in which individual agents only trust a limited number of agents they know individually can be described as a network.

From the above discussion, it is immediately clear that if trust is beneficial for economic efficiency, the loss from imperfect trust can be visualized as the difference between the actual trust network and the minimum network that would support all mutually beneficial trades. Following this reasoning, inefficiency is expected to be highest in societies where the trust network is very sparse (Granovetter (1995)). Inefficiency is also large when sub-groups who could benefit a lot from trading with each other are prevented from doing so by mutual isolation. This is true even if many links exist within each sub-group.

Social capital and public goods

In the preceding sub-section we discussed the role of trust in fostering exchange. Trust is also an essential ingredient in the delivery of public goods. In many cases, the state can organize the provision of public goods by taxing individuals. Whenever this is true, trust is not essential. But there are many forms of public goods that cannot be harnessed through state intervention.

In his work on PTA run schools, for instance, Coleman (1988) shows that parental involvement in school affairs has a beneficial external effect on student achievement, probably because it leads children to believe their parents care about their education. Parental involvement, in turn, requires trust to reduce and solve interpersonal conflicts and to minimize fears of free-riding. In this example, the externality is a public good that cannot be harnessed by state intervention. Voluntary participation by parents is essential.

In poor countries, there are many situations in which the state could, theoretically, intervene to provide a public good, but where it is unable to do so because its tax base

and its capacity to organize are limited. Collective action can serve as a substitute for the state. However, because it cannot rely on the coercive power of the state (e.g. the ability to tax and enforce contracts), collective action is much harder to set in motion. Two essential ingredients are then required: leadership and trust. A leader is required who is capable of convincing community members that they should voluntarily contribute to the public good. Trust is necessary to resolve conflicts among competing interests and to reduce fears of free-riding. Leaders can also help raise the level of trust in the community.

What the above discussion indicates is that delivering public goods via voluntary organizations depends critically on local trust and leadership. If these ingredients are absent, for instance after a civil war, then state intervention is likely to be much easier. Furthermore, good local leaders are rare. Projects that work well in one place because of strong local involvement need not be replicable elsewhere if local leaders are weak. Pilot projects of public good delivery through local communities may provide wrong signals if their placement is correlated with the presence of good local leaders who managed to attract the pilot project to their community.

II.iii Social capital and development

Much of the interest in social capital stems from the view that the absence of social capital represents one of the major impediments to economic development; Woolcock (1998) provides a wide ranging conceptual analysis of the role of social capital for developing societies and economies; a range of applications of social capital to economic development are collected in Dasgupta and Serageldin (2000) and Grootaert and van Bastelear (2002),. In fact, much of the current interest in social capital stems from the now classic book by Putnam, Leonardi and Nanetti (1993) which argues that Northern Italy developed faster than Southern Italy because the former was better endowed in social capital -- measured by membership in groups and clubs. One of the major claims in this literature is that social capital can facilitate the solution of collective action problems.

However, when focusing on advanced societies, the effects of social capital on economic performance are less obvious. For example, Putnam (2000), focusing on the U.S. experience since the 1950's, argues that social capital, defined as membership in formal and informal clubs, has declined monotonically since the 1950's. This is true for all states, all decades, and all measures of social capital. However, he finds no relationship between the speed of the decline of social capital and economic performance across U.S. states or across time periods. Further, the relationship between social capital and socioeconomic outcomes is even harder to characterize when one looks at subperiods. For example, the 1990's were a period of rapid economic growth in the U.S. yet it is also a period of rapid decline in social capital, at least based on the sorts of measures he uses. To be clear, Putnam does attempt to associate higher social capital with better socioeconomic outcomes, our point is that the relationship between the two for the United States is even at first glance relatively complicated.

The differences between the case of Italian regions and that of the United States is suggestive of how one might think about the relationship between development and social capital. One interpretation of these differences is that for the United States, generalized trust has improved over the period studied, so club membership has become less necessary.[5] In contrast, the Italian experience relates to an earlier period in which generalized trust may have been insufficient or incomplete and small clubs helped broaden the range of personalized trust. This raises the general possibility that clubs and networks are important at intermediate levels of development. Their function is to broaden the range and speed of social exchange beyond the confines of inter-personal trust. But once a sufficiently high level of generalized trust has been achieved, clubs and networks are no longer necessary and wither away (North (2001)). A similar kind of reasoning can be followed for public goods. In undeveloped economies, the state is weak

[5] In this discussion, we stipulate that Putnam's claims about declining U.S. social capital are correct. In fact, this claim has been subjected to important criticism. Skocpol (1996), has argued, for example, that while participation in local groups has declined, participations in larger organizations such as the American Association of Retired Persons has increased, and that what really needs be understood is the nature of voluntary group memberships and the like, rather than the number of memberships per se. See Skocpol (2003) for a detailed elaboration of this idea. One important implication of

and under-funded. Consequently it cannot organize the delivery of all needed public goods. This is particularly true for local public goods or for public goods that require a modicum of voluntary involvement to limit free-riding (of which corruption is but one manifestation).

Social capital provides an alternative. Clubs formed for non-economic purposes (e.g., religious worship) have leaders. In the absence of public good provision by the state, these leaders may decide to mobilize club members (e.g., the religious congregation) to provide missing public goods. History is replete with examples of faith-based organizations intervening to build schools and clinics and to provide a variety of public services. Here, sharing a common religious fervor is the basis for trust and the religious hierarchy provides the necessary leaders. Some large secular organizations have adopted similar practices – e.g., political parties yesterday, non-governmental organizations (NGOs) today.[6]

These issues have immediate implications for empirical work on social capital. The difficulty comes from the fact that first-best outcomes can in principle be achieved without paying attention to clubs and networks. Generalized trust in commercial contracts, for instance, can theoretically be achieved via laws and courts. Because of the possibility that revenues may be collectively raised via taxation, public goods can in principle be organized by the state at lower cost in terms of public mobilization and leadership skills. As North (1973,1990) has argued, the rise of the Western world is precisely due to the invention of institutions that protect property rights and make the state more effective at delivering public goods. Clubs, networks, and community-based voluntary organizations can improve efficiency in economic exchange and public good delivery. But these are typically second-best solutions. The first-best approach is generally to develop well-functioning legal institutions and state organizations.[7]

Skocpol's work for economists is that many of the measures that have been proposed to quantify social capital may be fundamentally flawed.

[6]One classic historical example is the role of the Social Democratic Party in organizing a range of social and cultural activities for its members in Imperial Germany, see Blackbourn (1997, chapter 8).

[7]Bowles and Gintis (2002) elaborate this type of reasoning, although in their view social capital plays a role in overcoming limits to government intervention generated by

Whether or not social capital raises efficiency we therefore argue depends on the level of institutional development. Suppose that laws and courts are insufficient to ensure respect of commercial contracts. This situation can arise anywhere (Bernstein (1996)) but it is probably most severe in poor countries where many transactions are small and buyers and sellers are too poor for court action to yield reparation (Bigsten *et al.* (2000), Fafchamps and Minten (2002)).[8] In such an environment, market exchange relies on a combination of personalized trust, legal institutions (e.g., to enforce large contracts and to punish thieves), and informal institutions (e.g., reputation sharing within business networks and communities). Whether or not social capital facilitates exchange can then be seen as a test of the strength and reach of formal institutions.

A similar line of reasoning holds for public goods. Public good delivery is best accomplished when the power of the state to tax and mobilize resources is combined with trust and community involvement. The reason is that, without voluntarily accepted discipline, government action is ineffective: taxes do not get paid, rules are not followed, civil servants become corrupt, and free riding reigns. Discipline in turn depends on the perceived legitimacy of government action and the degree of public involvement in the decision-making process. It also depends on identification with the political elites, sense of national urgency, and many other factors which are still poorly understood. The bottom-line, however, is clear: without some form of voluntary acceptance by the public, government efforts to provide public goods are likely to fail. Social capital is thus probably essential for public good delivery. But the forms it may take are likely to vary depending on local conditions, i.e., from generalized trust in government and formal institutions to interpersonal trust mobilized via clubs and networks.

II. iv. Social capital and equity

We have argued that trust is essential to both economic exchange and public good delivery. We have also argued that clubs and networks can facilitate search and provide

---

information constraints and so acts as a complement to government institutions in producing efficient outcomes.

[8]Except through forced labor, as in 19th century England and France. But this is now outlawed in most countries.

an imperfect substitute to generalized trust: in the absence of generalized trust, it may be necessary to rely on clubs and networks. Unlike generalized trust, however, clubs and networks often have distributional consequences that may be quite inequitable. The reason is that, unlike generalized trust, clubs and networks only offer a partial or uneven coverage of society. If the benefits of social capital principally accrue to network members, those who happen to be included benefit from increased efficiency but those that are excluded are penalized. As Fafchamps (2002) and Taylor (2000) have shown, the creation of clubs or networks can even penalize non-members. This is because members of a club or network find it easier to deal with each other and, as a result, may stop dealing with non-members.[9]

Clubs are least conducive to equity when membership is restricted to a specific group (e.g., men or whites) or when new members are not accepted (e.g., established firms only). Even when new members are accepted without restriction, historical events can shape the composition of clubs for decades whenever entry is slow. In this case, equal opportunity need not be realized because old members have enjoyed the benefits of membership for much longer. By extension, clubs are likely to have undesirable consequences on equity whenever (1) club membership is beneficial to members and (2) entry into the club is not instantaneous. Put differently, clubs raise equity concerns whenever they have real economic benefits.

The creation of clubs may thus reinforce polarization in society between the 'in' group and the 'out' group. Investing in social capital by promoting clubs can thus have serious equity repercussions. This is true even if we ignore the fact that certain clubs may collude to explicitly dominate or exclude others (e.g., the Ku Klux Klan). A similar situation arises with networks because better connected individuals profit from their contacts (Fafchamps and Minten (2002)). Social capital can be used by certain groups to overtake others, generating between-group inequality and political tension. To the extent that between-group inequality itself favors crime and riots and deters investment, promoting social capital by promoting specific groups may, in the long-run, be counterproductive.

[9] Of course, this is not to say that impersonal markets based on generalized trust treat all groups fairly. Statistical discrimination, for instance, naturally arises even in the absence

III. When does social capital matter?

The conceptual discussion has clarified the definition of social capital and its possible role in the development process. This discussion, however, has not precisely identified the conditions under which social capital matters. To achieve this, we need a general conceptual framework in which there is room for social capital to be beneficial.

III.i Sources of inefficiency

For social capital to increase Pareto efficiency, the decentralized equilibrium without social capital must not be Pareto efficient in the first place. Social capital can only have a beneficial effect in a second-best world. Deviations from first-best outcomes arise for a variety of reasons including externalities and free-riding, imperfect information and enforcement, imperfect competition, and the like. For social capital to be beneficial, it must therefore resolve or compensate for one of these sources of inefficiency. Secondly, whatever the source of inefficiency, there are only a limited number of ways by which social capital – or any other mechanism – may improve upon a decentralized equilibrium. First, it may resolve a coordination failure in an economy that has multiple Pareto-ranked equilibria. Second, it may alter individual incentives so as to replace the decentralized equilibrium with a superior one. Third, it may affect the technology of social exchange, for instance by opening new avenues for the circulation of information.

From these two preliminary observations, it is immediately obvious that social capital will never be the only possible solution to inefficiency. There always exist alternative mechanisms to solve coordination failure, improve individual incentives, and upgrade the technology of social exchange – such as contracts, vertical integration, state intervention, or redefinition of property rights. Of course, there are many circumstances in which social capital is a less expensive or simpler institutional solution, but it is important to recognize that it can never be the only one.

---

of clubs and networks (e.g. Fafchamps (2003)).

These observations have immediate implications regarding empirical investigation. Suppose social capital improves efficiency by solving a coordination failure problem. For this to occur, the economy must have multiple Pareto-ranked equilibria. Social capital provides the leadership or coordination device necessary to select a superior equilibrium among the many possible ones. Suppose further that the researchers have multiple observations of such economies, some with social capital and some without. Since nothing precludes these economies from achieving a high equilibrium without social capital, it is inherently difficult to test its effect. Furthermore, social capital may arise endogenously as an institutional response to an inferior equilibrium. To the extent that social capital does not always succeed in moving the economy to the better equilibrium, one could have the paradoxical situation in which economies with social capital are on average at a lower equilibrium than those without. This is a standard difficulty with multiple equilibria but it is not always adequately recognized in empirical work.

Even when there is a single equilibrium, social capital never is the only possible way of improving efficiency by altering incentives or technology. Identifying the effect of social capital requires that the researcher adequately control for other possible institutional solutions. Here too, self-selection is a concern.

III.ii Channels

The literature has identified a number of channels by which social capital improves efficiency. Most of these channels fall under one or a combination of the following three categories: information sharing, group identity, and explicit coordination.

Information sharing

It is a commonplace that human beings derive satisfaction from interacting with others. Socializing often involves the transfer of information, even if the purpose of socialization is not to transfer this information. The sharing of information is then a by-

product of social interaction, a Marshallian externality. To the extent that the shared information is economically useful, socialization generates a positive externality.

Socialization may also be initiated with the intent of acquiring a specific piece of information. In this case, the transfer of information is the purpose of socialization. Because interacting with others is also a consumption good, collecting information through socialization benefits from a kind of 'subsidy' relative to non-social forms of information collection (e.g., going to the library).

The literature on social capital contains many applications of this simple idea. Barr (2000), for instance, argues that social networks among Ghanaian entrepreneurs serve to channel information about new technology. Fafchamps and Minten (1999), Granovetter (1975,1995), Montgomery (1991), Rauch and Casella (2001) and many others have emphasized the role of business networks in conveying information about employment and market opportunities. Fafchamps (2004), Greif (1993), Johnson, McMillan and Woodruff (2000), Kandori (1992) and McMillan and Woodruff (2000) have brought to light the role of social networks in circulating information about breach of contract, thereby enabling business groups to penalize and exclude cheaters. Wade (1987, 1988) discusses the role of social capital in reducing incentive problems in teams by circulating information about effort. This point has also been made in the theoretical literature on industrial organizations, where the possibility for members of a team of workers to monitor and penalize each other has been shown to increase efficiency. Social capital may also circulate information about what tasks need to be done and when. Platteau and Seki (2002) provide an illustration of this idea in the case of Japanese fishermen and the coordination of their fishing efforts to minimize cost (e.g., exchange information about fish location) and maximize revenue (e.g., coordinate the landing of fish to maximize prices).

While the evidence provided is impressive, the literature remains somewhat naïve in its assumption regarding the ease with which accurate information can be exchanged. In practice, three conditions must be satisfied for social capital to raise Pareto efficiency through the sharing of information: (1) imperfect information must be the source of inefficiency; (2) there are disincentives to spread erroneous information; (3) there are no obstacles to Pareto efficiency other than imperfect information. Even if social capital

satisfies the first condition, it may not satisfy the other two. It is also important to recognize that the information sharing benefits generated by social capital can always be obtained in another way. For instance, information sharing can be explicitly organized and budgeted within a large organization, whether public or private (enterprise, NGO). To empirically test the effect of social capital, one should control for the possible presence of such organizations.

It is so customary to blame imperfect information for economic inefficiency that other sources of inefficiency, such as imperfect contract enforcement and insufficient protection of property rights, are sometimes disregarded. Fafchamps (2002), for instance, shows how the decentralized enforcement of contracts naturally takes the form of relational contracting, even without exchange of information. In this example, contract enforcement is the channel through which social capital raises efficiency, not information sharing. In his analysis of market institutions in sub-Saharan Africa, Fafchamps (2004) points out that incentives often exist to distort the conveyed information, either to hurt a competitor or to hide one's own shortcomings. Interviews with entrepreneurs suggest that gossip is never regarded as reliable information. Guaranteeing that accurate information is transferred through social networks requires the existence of punishment mechanisms – such as the loss of reputation – penalizing false reporting. Finally, there often are obstacles to Pareto efficiency other than imperfect information. The most common one is coordination failure. We revisit this issue below.

Group identity and modification of preferences

Under the general heading of group identity and modification of preferences, we put various effects that arise because identification with a group or network affects individual preferences and choices. Economists usually regard individual preferences as exogenously given and relatively stable over time. As psychologists have shown, however, individual preferences can be manipulated through advertising or propaganda. Individual preferences can also fluctuate over time in a systematic, somewhat predictable fashion. Impulses are one particularly relevant example of such phenomenon. Individuals

have been shown to violate their own stated preferences in response to an impulse – to eat, to drink, to buy.

This introduces time inconsistency in preferences. Because agents anticipate they may be subject to impulses, they often resort to various 'tricks' that limit their future choices – such as putting money on a savings account that cannot be accessed easily, or carrying a limited amount of cash when shopping. Agents may also voluntarily enter in restrictive social arrangements in order to protect themselves against their own impulses. Alcoholic Anonymous is a good example of such a process. Participation in Rotating Savings and Credit Associations (ROSCAs) can similarly be understood as a way of forcing oneself to save.

The literature on social capital is replete with descriptions of such virtuous processes. Because these descriptions implicitly assume that social capital alters individual preferences, they often seem alien to economists. One such claim often made in the literature is the idea that social capital favors altruism and raises concerns for the common good – the 'touchy-feely' side of social capital. To see how even a minor increase in altruism can raise efficiency, consider a standard Prisoner's Dilemma (PD) game with standardized payoff matrix:

	Cooperate	Defect
Cooperate	(1,1)	(*-a*,*b*)
Defect	(*b*,*-a*)	(0,0)

with $a > 0$, $b > 1$. It is standard that (Defect,Defect) is the unique Nash equilibrium. Now suppose that players become altruistic, so that their utility is the weighted sum of their individual payoff Π_i and their opponent's individual payoff Π_j, so that $U_i = (1-\alpha)\Pi_i + \alpha\Pi_j$ where $\alpha > 0$. In this case, Defect is no longer necessarily a best response strategy; (Cooperate,Cooperate) is now a Nash equilibrium if $1 > b(1-\alpha) - a\alpha$ or equivalently, $\alpha > \frac{(b-1)}{b+a}$. This condition can be satisfied for values of α well below one half, implying that, depending on the values of a and b, even moderate levels of

altruism can eliminate the Prisoner's Dilemma. Similar reasoning can be applied to games with inferior equilibria, such as the assurance game: in these games some altruism can also eliminate Pareto inferior outcomes. The intuition behind this result is obvious: the more players internalize others' payoffs, the more they care about Pareto efficiency. When both players give equal weight to their payoff and others', they only care about aggregate welfare, what we call the common good. In this case, the equilibrium is always Pareto efficient.[10] Altruism provides an efficient solution to free-riding – a principle that most religions seem to have discovered centuries ago.

The relationship between altruism and social capital probably has to do with group identity (Akerlof and Kranton (2000)). Economic experiments using the dictator game and the trust game indeed suggest that agents exhibit more altruism and play more cooperatively if they have been induced to identify with a group (e.g., Fershtman and Gneezy (2001)).[11] This is true even if members of the group are unknown and even if they are not even seen during the experiment. These results suggest that group identification may trigger agents to adopt more altruistic preferences, thereby yielding more efficient group outcomes. If identification with a group is necessary for preferences to become more altruistic and better aligned with the common good, efforts to foster a sense of community may naturally be seen as an essential component of social capital by many researchers. This probably explains why community building is often construed as a way to foster social capital.

Social capital may also affect preferences in other ways. As argued by Fafchamps (1996) and Platteau (1994a), several mechanisms can be used to enforce contractual obligations: legal and extra-legal penalties, loss of reputation, and guilt. These same mechanisms can enforce contributions to the public good in case individual preferences are not aligned with the common good. By circulating information, social capital can magnify reputational sanctions, a point we have discussed in the previous sub-section. Group identification can also raise guilt for acting against the group's common interest.

---

[10] Note that the common good equilibrium is Pareto efficient in both the original, selfish preferences Π_i and in the altruistic preferences $U_i = (1-\alpha)\Pi_i + \alpha\Pi_j$.

[11] In the trust game players play sequentially. Player 1 gives an amount X to player 2. This amount is multiplied by the researcher, usually by 2 or 3. Player 2 then gives an amount Y to player 1. There is no repetition.

In our PD game, this is formally equivalent to deducting the subjective cost associated with guilt, call it g, from the payoff b associated with defection. If this feeling is strong enough so that $b-g<1$, defection is deterred. Since Max Weber, the literature on market development has emphasized the role played by religion in fostering business honesty (Ensminger (1992), Geertz, Geertz and Rosen (1979), Poewe (1989)). Communist work ethics propaganda can be seen as a similar effort to improve team performance by raising guilt among shirkers.

By favoring identification with a group, social capital may also affect preferences through mimicry. In the literature, this idea appears in many guises, the phrase most commonly used being 'role model'. Coleman's example of PTA-run schools is a good illustration. According to Coleman, children whose parents are involved in running the school adopt a more positive attitude towards study. This change in preferences cannot be understood as altruism: it is in the children's long-term self-interest to study. Nor does it appear to be purely the result of a sharpened sense of guilt for not studying. Rather it is related to a demonstration or role model effect: children change their preferences to mimic that of their parents. By visibly and credibly demonstrating their positive attitude towards school, parents induce a change in attitude among their children.

This kind of phenomenon is related to what economists have called 'herding behavior', that is, the drive to mimic the behavior of others. More research is needed in this area to fully comprehend the phenomenon and its implications for economic efficiency. As has been argued formally in Blume (2002), however, mimicry need not result in superior equilibria: nothing in mimicry itself precludes agents from copying bad behaviors instead of good ones. One famous example is that of a group of high school students who refused to take their graduation exam as a symbol of group identity, even though doing so hurt them all. Other examples of bad mimicry involve hazing, gang rape, crime culture, and the like. Unlike altruism, mimicry is a double-edged sword.

Coordination and leadership

Some of the beneficial effects of social capital on preferences occur by osmosis, without any purposeful action by anyone: people chat around a glass of beer and, quite by

chance, a relevant piece of information is exchanged. In many cases, however, the benefits of social capital are only achieved through purposeful action: someone has to want to improve the group's welfare and must do something about it for benefits to materialize. This is particularly true of any benefit that requires coordination in order to be achieved.

This raises a host of difficult issues having to do with the decision making process within groups. It is well beyond the scope of this Chapter to discuss these issues in detail. A few remarks are nevertheless in order. First, two essential ingredients seem to play fundamental roles in purposeful group action: leadership, and rules regarding group decision making. At this level of generality, their respective role is unclear. What is inescapable, however, is that neither of them constitutes social capital.

In very informal groupings, leadership is likely to be essential to alter individual preferences and elicit voluntary contributions to the common good. While social capital may assist the action of leaders by facilitating the circulation of information and favoring group identification, the respective roles of leadership quality and social capital are likely to be extremely difficult to disentangle. This has important implications for empirical work: if good leadership is required to achieve the coordination required to benefit from social capital, testing the effect of social capital requires controlling for the quality of leadership.

This observation also has implications for policy. Good leaders may improve efficiency by using the levers of social capital – e.g., by fostering altruistic preferences and concern for the common good; favoring group identification; preaching good behavior and making free-riders feel guilty; encouraging mimicry of good behavior through role models and the manipulation of group symbols and representations (e.g., religion, ideology). This is what practitioners in the field call 'building social capital'.[12] Many NGOs, for instance, are engaged in precisely this kind of work. Sometimes they focus on the identification and training of local leaders, something to which many NGOs refer as an example of 'capacity building' (Barr, Fafchamps, and Owens (2004)).

[12]To a number of economists, these forms of policy intervention may seem unusual because they have no effect on material incentives but operate only through mental representations. We revisit these issues in greater detail below.

Purposeful coordination can also be obtained through formal rules by which decisions are made and deviance penalized. A simple majority rule combined with fines and jail sentences for free-riders is in many cases sufficient to reach efficiency. As long as free-riding is not so prevalent as to overwhelm policing, punishments directly alter incentives in ways that align individual behavior with the common good. In this case, social capital plays little role – except perhaps in coordinating not to overwhelm the enforcement apparatus. Leadership also becomes less critical since there is no need for a charismatic leader who can affect individual preferences directly. All that is required is a 'bureaucratic' leader who can apply and enforce the rules decided by the group.

A proper investigation of the importance of social capital in economic life therefore requires a careful analysis of the rules by which decisions are reached. It is important not to credit social capital with outcomes due to formal rules. This means distinguishing between the benefits resulting directly from formal organization and the indirect benefits members derive from contact with each other. For instance, the Rotary Club has a decision making body to coordinate the date and venue of its next dinner. The coordination benefit of meeting on the same day in the same place follows directly from the Club's formal rules. But once at the dinner, there is probably no coordinated mechanism to share information among members.

This same sort of reasoning applies to schools. In addition to the effects of student attitudes discussed by Coleman, PTA-run schools have an organizational structure different from that of other schools. In particular, decisions are taken differently and funding is allocated in a different manner when parents and teachers possess decisionmaking power in schools. As Jimenez and Sawada (1999) have shown in the case of El Salvador, PTA-run schools tend to provide greater remuneration and select better teachers than other schools. These schools also exhibit lower rates of teacher absenteeism. At least part of these differences may plausibly be attributed to differences in funding and internal decision-making rules. Disentangling these effects from those of social capital is likely to be difficult and contentious.

III.iii. Formal theory

While the ideas associated with social capital have been linked to many strands of modern microeconomic theory, there has been relatively little formal modeling of social capital per se. One reason for this, we conjecture, is the absence of a generally accepted and coherent definition of social capital, as discussed.

In terms of the efforts to embody social capital in formal economic models, one approach that has been taken is to incorporate social capital in models in the context of repeated prisoner's dilemma games. In environments in which agents change partners, the sustainability of a cooperative equilibrium depends on either the likelihood with which a match today will be repeated in the future and/or the ability of an agent to access information about the past behavior of a new partner (Kandori (1992)). In this context, social capital is interpreted in terms of the factors that facilitate the existence of a cooperative equilibrium. Routledge and von Amsberg (2003), using a prisoner's dilemma environment of the type we described above, define social capital as present whenever a cooperative equilibrium exists; the key variable that determines whether cooperation can occur is the probability of trade between a pair of agents. Intuitively, if this probability is high, two agents meeting today are likely to meet in the future, so that any loss from cooperation today is compensated by future cooperation in the repeated relationship. Routledge and von Amsberg apply this idea to study how migration across regions or sectors, can, by lowering the likelihood of repeated interactions, lead to a loss of social capital. Annen (2003) defines social capital as an individual's reputation for cooperation in prisoner's dilemma games. In his analysis, this reputation depends on the extent to which information transmission about past behavior is reliable and the complexity of the network in which agents interact. Changes in either reliability or complexity can thus alter levels of social capital. Annen focuses on the question of when increases in network complexity lead to a reduction of network size or an increase in network size accompanied by greater investment in communication capacity.

Other formal theory relevant to social capital includes efforts to model the notions of trust and trustworthiness. Zak and Knack (2001) study a general equilibrium growth model in which agents facing moral hazard problems decide how much to invest in monitoring. The presence and strength of formal and informal sanctions for dishonesty are shown to have powerful implications for growth because of their role in reducing the

need to invest in monitoring. Another approach to modeling trust is due to Somanathan and Rubin (2004), who study the evolutionary stability of honest types in a population.

Perhaps the most important contribution to formal theory is Dasgupta (2002) which provides a wide ranging discussion of the relationship between social capital and formal modeling. Dasgupta argues that social capital should not be defined in terms of the presence of cooperation or some other outcome; rather that it should be regarded directly as social structure.

> "...social capital is most usefully viewed as a system of interpersonal networks...If the externalities network formation gives to are "confined", social capital is an aspect of "human capital", in the sense economists use the latter term. However, if network externalities are more in the nature of public goods, social capital is a component of what economists call "total factor productivity." (pg. 6-7)

Dasgupta's analysis is important as it indicates how the role of social capital in growth cannot be reduced to the addition of a variable to a linear cross-country growth regression. His analysis is also important in its recognition that theoretical claims about the desirability of the sorts of social structures that have been equated to social capital are to some extent artifices of particular modeling assumptions. For example, he argues that the claim that repetition of a one-shot game necessarily benefits the players of the game is not a generic finding and in fact does not generally hold for payoff structures other than the prisoner's dilemma, going on to argue that work such as Fudenburg and Maskin (1996) shows how social capital can lead to exploitive relationships. As such Dasgupta's analysis makes clear how functional notions of social capital are inconsistent with rigorous theorizing. Other conceptual discussions of social capital and social science include Ostrom and Ahn (2002) and Paldam and Svendsen (2000); the former is particularly interesting to contrast with Dasgupta (2002) as it is written from the perspective of non-economists and indicates some of the conceptual gaps between economists and other social scientists on this topic.

IV. From theory to empirics: econometrics and social capital

Having clarified the relationship between social capital and the efficiency of social exchange, we now turn to the statistical analysis of the effects of social capital. We first revisit the points raised in this section, such as the distinction between individual and aggregate efficiency effects. We then ask whether it is possible to uncover social capital effects from the sorts of data available to social scientists. In particular, we discuss the issue of identification, that is, of whether a role for social capital can be uncovered when other types of social effects may be present.

Standard practice in economics and sociology is to run regressions of some outcome of interest against a set of controls and some asserted empirical proxies for social capital. These regressions are often justified by an informal argument that the empirical proxies act as instrumental variables for the unobserved "true" social capital measure. At one extreme, one finds analyses such as Furstenburg and Hughes (1995) in which the probability that an individual drops out of school is related to variables such as the presence of a father in the household or the educational aspirations of the person's friends. In contrast, studies such as Knack and Keefer (1997) attempt to explain growth differences across entire countries using survey measures of trust.

In this section, we discuss some general econometric issues that arise in social capital studies of this type. We first examine difficulties inherent in the estimation of the benefits from social capital on the basis of individual data. These difficulties are not specific to social capital and are shared by other externalities. But they are often ignored in empirical work.

Second, we discuss the question of model specification. In particular, we review some requirements for treating a given social capital regression as causal. Next, we discuss identification. In this case, we assume that a researcher has the "correct" model of some outcome of interest and ask whether observational data on the phenomena will allow for the identification of a causal relationship between social capital and the outcome.

The basic econometric issues associated with identifying a role for social capital may be understood in the context of cross-sections. While panel data have certain advantages, notably that they allow for the researcher to control for fixed effects across units, the conditions under which social capital effects may be identified are not qualitatively different.

IV.i. Externalities and individual vs. aggregate effects

As we have discussed in Section II, the literature on social capital is interested in externalities arising from coordination failure. Much of the empirical work on social capital seeks to identify the effect of social capital on an outcome variable of interest, say, ω_i. This variable of interest can be measured at the aggregate level – e.g., country growth – or at the individual level – e.g., performance of a pupil on an exam. Empirical work on social capital can thus be divided into individual and aggregate level regressions.

The first difficulty many researchers encounter is that individual returns to social capital often are poor predictors of aggregate externalities. There are two main reasons for this: fallacy of composition and free riding. A fallacy of composition arises whenever social capital pegs individuals against each other. In a situation of competition for a finite resource, the gains made by those with more social capital lead to losses for those without, relative to a situation without social capital. Free riding is the opposite situation in which aggregate social gains are larger than those appropriated by the owners of social capital. We discuss them in turn.

Fallacy of composition

To illustrate fallacy of composition, consider a simple job search example inspired by Granovetter's work. Suppose there are M job openings and N job seekers, all identical, with $N>M$. Suppose that employers and workers do not know each other and are matched at random. Since $N>M$, all positions are filled and each worker has an equal probability of getting a job $\frac{M}{N}$. Total surplus is the sum of employer and worker

surplus. Since all workers are equivalent, total surplus is the same irrespective of which workers get the available jobs.

Next suppose that, because of interpersonal connections, a group of workers C hears about the open positions before other workers. Further suppose that $C < M$. Consequently C workers get a job with probability 1. Other workers get the remaining jobs with probability $\frac{M-C}{N-C}$ which is smaller than $\frac{M}{N}$. Total surplus is unchanged since workers are equivalent. Social networks – in this case the existence of a better connected group of workers – have no effect on the efficiency of social exchange. But they have important distributional consequences, which can be measured by regressing the probability of obtaining a job on group membership. Doing so in our example would yield a coefficient of $1-\frac{M-C}{N-C}$ on membership in the group even though the net effect of social networks on aggregate welfare is zero. What this example illustrates is that social networks can have private returns even when they have no effect – other than distributional – on the efficiency of social exchange. Observing private returns to social networks should therefore not be construed as evidence of social capital. In our example, social networks actually generate a discriminatory outcome, which is inconsistent with equality of opportunity as conceptualized by Roemer (1998) for example.[13]

The above reasoning can be extended to situations where groups, not individuals, compete with each other. Consider, for instance, high schools competing to place their graduates at Harvard. We assume that the number of admissions in Harvard is fixed and that the university selects the students with the best grades on a standardized test. Suppose that Coleman is right and that, because of the social capital effects of parental involvement in school affairs, students in PTA-run schools obtain better grades. As a result, they are more likely to go to Harvard than students from non-PTA schools. Whether or not this raises social welfare depends on how critical high school education is to university learning.

[13] A similar example could be constructed in which it is the effect of social capital on trust that matters. For instance, imagine silk produced in China and consumed in Europe. Chinese silk producers do not trust European consumers so that direct sale is not possible.

To illustrate this point, suppose that students learn all they need to know at Harvard. The only purpose of high school education is to screen out less able students. Further assume that the minimum grade required to be admitted at Harvard is higher than the grade necessary to earn one's degree: some applicants do not get in even though, if they did, they would earn their degree. In this case, the role of social capital is again to enable one group – students in PTA schools – preferential access to a rationed resource – admission at Harvard. The effect of social capital is distributional. Regressing the probability of admission in Harvard on social capital would yield a positive coefficient even though, in this example, the effect of social capital on the efficiency of social exchange is zero. Of course, we do not claim that the above example is an accurate depiction of the education system. The only purpose of the example is to illustrate the danger of estimating the beneficial effect of social capital by comparing individual or group outcomes according to whether or not they have social capital. Whenever social capital enables one group to displace another, a statistical comparison of the two groups is bound to overestimate the efficiency gain from social capital.

This example exposes another ambiguity of the concept of social capital. In our review of definitions of social capital, we noted that most authors associate social capital with the idea of beneficial group externalities. In the above – admittedly extreme – example, groups of students in PTA-run schools benefit from the social capital generated by their parents. But society as a whole does not. According to our definition, there is social capital at the level of each group but not at the aggregate level. This contradiction serves to remind us that it is perilous to define a social process as necessarily having beneficial effects.

Free riding

It is also possible that social capital generates beneficial externalities but yields no (or few) individual returns for the holders of social capital. A case in point is when the external effects of social capital are fully captured by outsiders – i.e. individuals or

---

A group of traders who manages to gain the trust of both producers and consumers can then capture the silk trade.

groups who are outside the social networks or do not share the norms and values of the group – who do not incur the cost of generating the externality.

To see this, consider N groups of fishermen tapping the same fishing ground.[14] Without collective action, there is over-fishing. Suppose that fishing groups with better social capital enforce self-restraint – either through shared norms or through relational contracting – while others do not. Gains from self-restraint are shared among all fishermen, irrespective of whether they have social capital or not. Social capital increases aggregate social welfare but fishermen with less social capital have higher profit because they free ride: they benefit from the self-restraint of others without having to incur any cost. Regressing fish catch on social capital would result in a zero or negative coefficient on social capital even though it has a positive social return.

The externality can also be pecuniary. Keeping the fishing example, a similar result obtains if the fishing groups do not share a common fishing ground but sell their fish on the same market: social capital makes collusion to restrict supply possible since all fishermen benefit from higher fish prices.[15] To ascertain the effect of social capital, one needs to compare fishing groups who do not compete with each other by either accessing the same fishing ground or by selling fish on the same market.

What these examples demonstrate is that, in the presence of fallacy of composition or free riding, individual returns from social capital are poor indicators of aggregate returns. If social capital enables certain individuals or groups to capture rents at the expense of others (e.g., jobs in a non-clearing labor market, entry at Harvard when the entry criterion is excessive), individual returns to social capital exceed social returns, and social capital generates unequal outcomes. In contrast, if social capital generates positive externalities not fully appropriated by owners of social capital, individual returns underestimate social returns.

IV.ii. Model specification

---

[14] This example is inspired by the work of Platteau and Seki (2002) on Japanese fishermen.

[15] An example of this situation is OPEC: not all oil producing countries are member, but they all benefit from higher prices even though only members of the cartel restrict their production.

Exchangeability

As we have noted, social capital studies have been applied to a remarkably large number of units of observation, ranging from individual farmers to countries. One natural question is whether these studies in fact use comparable observations. At an abstract level, comparability of observations is a requirement for virtually all causal studies. We raise the question in the context of social capital studies for several reasons.

First, social capital studies, particularly those that employ aggregate data, often use relatively crude sets of control variables. As a result, the residuals in the sample will contain forms of heterogeneity that call into question the placement of the observations in a common regression.

Second, social capital studies often fail to account for the reasons why different agents come to have different levels of social capital. As Durlauf (2002c) states

> ...statistical analyses of social capital typically compare outcomes for individuals or aggregates who have social capital versus those who do not. These studies, in turn, typically do not incorporate a separate theory of the determinants of social capital formation, although they do often employ instrumental variables to account for the endogeneity of social capital. However, without a theory as to why one observes differences in social capital formation, one cannot have much confidence that unobserved heterogeneity is absent in the samples under study. (pg. 464)

Notice that this argument is more general than simply arguing that social capital is an endogenous variable. Since the groups in which individuals are organized often are endogenous, there will be various forms of sample selection that need to be accounted for in empirical work.

To see that these are more than abstract concerns, consider the regressions employed in Helliwell and Putnam (2000) to show the effects of social capital on economic growth. These authors regress regional output growth in Italy against initial output and measures of civic community, institutional performance, and citizen satisfaction. They find that these three measures explain persistent differences in regional growth rates and conclude that this supports social capital explanations of

economic performance. Among the many questionable assumptions that underlie such a conclusion is the assumption that the regression they employ is using comparable objects as observations. In other words, the analysis assumes that each observation is generated by a common growth process. What must be assumed about the growth process in different regions when one includes Northern and Southern Italian regions in a regression? One answer to this question is that one must assume that given the variables included in the regression, the errors for the observations of different regions cannot be distinguished, at least from the perspective of their distributions. Put differently, one must assume that the regression is such that there is no reason to expect that the error from a particular region has a nonzero expected value, for example. But how can a regression of this crudity make such a breathtaking claim? The historical and social science literatures give any number of reasons why this assumption is false in contexts such as Italian regimes. But if the assumption is false then one cannot defend the interpretation provided by Helliwell and Putnam (2000) for their regression results.

Brock and Durlauf (2001b) argue that a way to formalize the notion of comparability is via the mathematical concept of exchangeability. We introduce this formalism as it provides a way of providing a link between the ways one thinks about data as a social scientist and the sorts of statistical assumptions that underlie regression exercises.

Suppose that for each of I observations, one has associated information F_i. This information may include factors that are quantifiable, such as the savings rate of a country, as well as factors that are not necessarily quantifiable, such as knowledge of a country's culture. Suppose that some outcome ω_i is generated by the linear model

$$\omega_i = \gamma Z_i + \eta_i \tag{1}$$

where Z_i represents that part of F_i that is controlled for in the regression. Typically, models such as (1) are interpreted as meaning that, except for differences in the value of Z_i, ω_i may be thought of as draws from a common distribution, which in turn means that the η_i's are drawn from a common distribution. Notice, however, that this notion of

being drawn from a common distribution should be determined relative to the complete information set available for each observation, i.e. F_i. Hence, interpretation of (1) presupposes that having controlled for the various Z_i's, one has no information that allows one to distinguish the residuals. Formally, the errors η_i are F_i -conditionally exchangeable, which means that

$$\mu\left(\eta_1 = a_1,\ldots,\eta_K = a_K \left| F_1 \ldots F_I \right.\right) = \mu\left(\eta_{\rho(1)} = a_1,\ldots,\eta_{\rho(K)} = a_K \left| F_1 \ldots F_I \right.\right) \qquad (2)$$

where $\rho(\cdot)$ is an operator that permutes the K indices.

Exchangeability is a useful formalization because it creates a benchmark for the assessment of empirical studies. In fact, many of the standard problems that arise in regression analysis amount to exchangeability violations. For example, when a regressor is omitted from a regression, this will mean that the errors in (1) are no longer exchangeable as the distribution of a given error will depend on the distribution of the included and omitted variables. Similarly, if there is parameter heterogeneity between observations, this will imply that the distribution of a given error depends on which country it is associated with. To take a third example, self-selection can induce exchangeability violations as the errors associated with one observation may be differentiated from other differences in the implications of self-selection for the conditional expectations of the residuals. To be clear, as Brock and Durlauf (2001b) observe, exchangeability is not necessary for causally interpreting regressions. For example, heteroskedasticity in errors is an exchangeability violation, but is compatible with a structural regression interpretation. What we argue here is that good empirical practice requires that one assess whether conditional exchangeability of errors holds for the regression under study. To be more precise, we believe that a good empirical practice is to ask, for a given regression specification whether, given the information a researcher possesses about the individual observations, the researcher can justify the assumption of (2) and if not, determine whether the regression retains the interpretation the researcher wishes to place upon it.

Instrumental variables

As observed above, in many contexts social capital is endogenous social capital. The problem of endogeneity is obvious in many contexts; when one talks about membership in organizations, one must account for the fact that membership is a choice variable. In other cases, the endogeneity problem is more subtle. Measures of trust are often used to characterize social capital. Since trust presumably is related to trustworthiness in actual behavior, such measures will exhibit endogeneity problems as well.

Many researchers have recognized that social capital is endogenous and so have employed instrumental variables to allow for consistent estimation of parameters. Leaving aside issues of self-selection that are not often not appropriately addressed by instrumental variables approaches, the use of instrumental variables in social capital studies can be subjected to criticism. Specifically, in many social capital studies the choice of instrumental variables often appears to rely on ad hoc and untenable exogeneity assumptions.

For example, Narayan and Pritchett (1999), using village level data, argue that measures of village level trust can instrument for measures of group memberships. In their analysis social capital effects are argued to occur when one individual's "associational life" affects others in his village; measures of associational life include factors such as the number of group memberships. Since associational life may be a consumption good and thereby an increasing function of individual income, Narayan and Pritchett argue that it must be instrumented if one wants to identify how social capital causally affects income. Yet, there is little reason that such a variable is a valid instrument. As pointed out above, if trust is related to trustworthiness, as presumably is the case, then there is no reason why trustworthy behavior is any different than membership in an organization in terms of whether it is a choice variable. And without a theory of what determines trustworthy behavior, there is little hope of identifying credible instrumental variables for it in these types of regressions.

The choice of instrumental variables is often one of the most difficult problems in empirical work. In social capital contexts, the absence of explicit modeling of the

process by which groups are formed and social capital created means that an empirical researcher is forced to rely on intuition and guesswork. While this does not condemn all studies using instrumental variables, we do believe that inadequate attention has been paid to justifying instrumental variables in social capital contexts.

Group effects versus social capital effects

A final specification issue in social capital studies concerns the question of distinguishing between social capital and other group effects. There is no shortage of reasons why group memberships influence individuals. For example, in recent models of income inequality, primary emphasis has been given to peer group effects and role model effects as influencing educational outcomes for youths. This creates a relationship between the outcomes for a given youth and the outcomes of others in his community of residence.[16] In many modern growth models, a key assumption is the presence of various types of increasing returns to scale that are produced by externalities. These types of models often take the form of positing that the productivity of a given actor depends on the human and physical capital stocks of others. From the perspective of statistical modeling, the description of individual behavior will require the incorporation of various group-level variables.

From the perspective of empirical work, the problem is simple. If one claims that a social capital effect is present for some behavior on the basis of the statistical significance of a group-level variable, this claim will not be credible unless one is able to argue that the group-level variable is capturing social capital versus some alternative group-level effect. This problem is particularly serious when social capital is endogenous, since aggregate levels of social capital are then determined by other group-level variables, which, in absence of strong prior information, presumably include whatever aggregate variables have been omitted from a regression explaining outcomes.

[16]See Durlauf (2001,2002a) for discussion of a range of possible group-level influences on individual behavior.

IV.iii. Identification

The question of social capital and other group effects leads to the question of identification. In this section, we assume that the model under study is correctly specified and evaluate what model parameters can be recovered from observational data. This work is developed in Durlauf (2002c), a paper which builds on early work by Manski (1993) and later work by Brock and Durlauf (2001a,c) on identifying group effects in data. Our basic framework treats the level of social capital in a community as an endogenous variable that represents the aggregation of individual-specific social capital levels (for example, investments in individual-specific social capital as in Glaeser, Laibson, and Sacerdote (2002)). As such, the determination of how social capital effects individuals is an example of the "reflection problem" that Manski's seminal (1993) paper characterizes; identification problems arise when one needs to distinguish the effects of the choices of others versus the characteristics of others on an individual. Identification questions when social capital is exogenous are discussed separately.

IV.iii.a. Individual-level Data

We first consider the case where one wishes to understand the effect of social capital on some individual outcome ω_i. For individual-level data, linear versions of social capital models can be expressed as follows. Suppose that each agent i is a member of some group $g(i)$. Each individual chooses an outcome variable ω_i that is linearly dependent on some control variables. Assume these variables are of four types: an r-dimension vector of variables that are measured at the individual level, X_i; an s-dimension vector of variables (often called contextual effects) that are measured at the group level and are predetermined at the time that choices are made, $Y_{g(i)}$; an individual's expectation of the average choice of others, $\mathrm{E}\left(\omega_{g(i)} \middle| F_{g(i)}\right)$ (called an endogenous effect, cf. Manski (1993)), where this expectation is made conditional on some information set $F_{g(i)}$; and expected social capital in the community, $\mathrm{E}\left(SC_{g(i)} \middle| F_{g(i)}\right)$. The assumption that

individual behavior depends on expected rather than actual social capital does not result in any loss of generality. Similarly, our assumption that agents react to the expected behaviors and social capital levels in their group rather than the expected levels among group members other than themselves has no bearing on the analysis, cf. Brock and Durlauf (2001a,c).

We assume that the X_i and $Y_{g(i)}$ vectors are components of the information sets from which expectations are formed; these expectations are further assumed to be rational, so we work with mathematical expectations rather than subjective beliefs. The behavioral outcome is described by

$$\omega_i = k + cX_i + dY_{g(i)} + J_1 \mathrm{E}\left(\omega_{g(i)} \middle| F_{g(i)}\right) + J_2 \mathrm{E}\left(SC_{g(i)} \middle| F_{g(i)}\right) + \varepsilon_i \tag{3}$$

In order to close the model, it is necessary to specify how group level social capital is determined. We assume that group level social capital is the average of individual social capital levels, SC_i. These levels are determined by an individual-level behavioral equation that is analogous to (3):

$$SC_i = \bar{k} + \bar{c}X_i + \bar{d}Y_{g(i)} + \bar{J}_1 \mathrm{E}\left(\omega_{g(i)} \middle| F_{g(i)}\right) + \bar{J}_2 \mathrm{E}\left(SC_{g(i)} \middle| F_{g(i)}\right) + \eta_i \tag{4}$$

The identification problem amounts to asking whether the parameters in (3) are uniquely determined by the reduced form equations that describe ω_i and SC_i. In order to solve for these reduced form equations, one first applies an expectations operator to both sides of (3) and (4). For the outcome equation,

$$\mathrm{E}\left(\omega_{g(i)} \middle| F_{g(i)}\right) = k + cX_{g(i)} + dY_{g(i)} + J_1 \mathrm{E}\left(\omega_{g(i)} \middle| F_{g(i)}\right) + J_2 \mathrm{E}\left(SC_{g(i)} \middle| F_{g(i)}\right)$$

or

$$\mathrm{E}\left(\omega_{g(i)}\middle|F_{g(i)}\right)=\frac{k+cX_{g(i)}+dY_{g(i)}+J_2\mathrm{E}\left(SC_{g(i)}\middle|F_{g(i)}\right)}{1-J_1} \tag{5}$$

and for the social capital equation

$$\mathrm{E}\left(SC_{g(i)}\middle|F_{g(i)}\right)=\bar{k}+\bar{c}X_{g(i)}+\bar{d}Y_{g(i)}+\bar{J}_1\mathrm{E}\left(\omega_{g(i)}\middle|F_{g(i)}\right)+\bar{J}_2\mathrm{E}\left(SC_{g(i)}\middle|F_{g(i)}\right)$$

or

$$\mathrm{E}\left(SC_{g(i)}\middle|F_{g(i)}\right)=\frac{\bar{k}+\bar{c}X_{g(i)}+\bar{d}Y_{g(i)}+\bar{J}_1\mathrm{E}\left(\omega_{g(i)}\middle|F_{g(i)}\right)}{1-\bar{J}_2} \tag{6}$$

In these expressions, $X_{g(i)}$ is the within-group average of X_i and represents the relevant set of variables that relate individual characteristics of group members to the group-level behaviors. Subsituting out $\mathrm{E}\left(\omega_{g(i)}\middle|F_{g(i)}\right)$ and $\mathrm{E}\left(SC_{g(i)}\middle|F_{g(i)}\right)$ in (3) and (4) using the expressions in (5) and (6) produces reduced form expressions for ω_i and SC_i. Durlauf (2002c) verifies the following proposition, which describes necessary conditions for identification.

Proposition 1. Identification in linear individual-level models with social capital

Identification of the parameters in eq. (3) requires

i. The dimension of the linear space spanned by elements of $\left(1, X_i, Y_{g(i)}\right)$ is $r+s+1$.

ii. The dimension of the linear space spanned by the elements of $\left(1, X_i, X_{g(i)}, Y_{g(i)}\right)$ is at least $r+s+3$.

What this proposition states is that identification depends critically on the relationship between the vector $X_{g(i)}$ that does not appear in the behavioral equations (3) and (4) and the vectors X_i and $Y_{g(i)}$ that do appear in these equations. Intuitively, the key idea is that identification of equation (3) fails if $\mathrm{E}\left(\omega_{g(i)}\middle|F_{g(i)}\right)$ and $\mathrm{E}\left(SC_{g(i)}\middle|F_{g(i)}\right)$ are linearly dependent on the other terms in the regression, i.e. $\left(1, X_i, Y_{g(i)}\right)$. Each of these variables is a linear function of $Y_{g(i)}$ and $X_{g(i)}$. So, if $X_{g(i)}$ is linearly independent of these other regressors, identification may hold.

What does this theorem require in terms of empirical implementation? A key requirement is that there are at least two X_i variables whose within-group averages are not elements of $Y_{g(i)}$. The existence of such variables will of course depend on context. For example, one can imagine situations in which an individual's age affects his behavior, but not the average age of others in his group. The need for such prior information illustrates how field work and qualitative studies can augment formal statistical analyses.

IV.iii.b. Aggregate data

A number of social capital studies employ data that are aggregated. Typically, these studies explore the average behavior of groupings which define the social environment for the individuals that comprise them. From the perspective of estimation, one can think of such models as taking within group averages of (3) and (4), so that

$$\omega_g = k + dY_g + J_1\mathrm{E}\left(\omega_g\middle|F_g\right) + J_2\mathrm{E}\left(SC_g\middle|F_g\right) + \varepsilon_g \tag{7}$$

and

$$SC_g = \overline{k} + \overline{d}Y_g + \overline{J}_1\mathrm{E}\left(\omega_g\middle|F_g\right) + \overline{J}_2\mathrm{E}\left(SC_g\middle|F_g\right) + \eta_g \tag{8}$$

where ω_g and SC_g are group level averages.

Necessary conditions for identification in this case are also developed in Durlauf (2002c). To characterize these conditions, let $H_{\omega,g}$ and $H_{SC,g}$ denote the linear spaces spanned by those regressors Y_g with nonzero coefficients in equations (7) and (8) respectively. Let $H^c_{SC,g}$ denote that part of $H_{SC,g}$ that is orthogonal to $H_{\omega,g}$ (i.e. the linear space formed by the orthogonal complements of any basis of $H_{SC,g}$ after being projected on $H_{\omega,g}$). These spaces are used in the following proposition on identification.

Proposition 2. Identification of social capital effects with aggregate data

i. Identification of the parameters in eq. (7) requires that the dimension of the linear space $H^c_{SC,g}$ is at least 2.

ii. If J_1 is known to equal 0, then identification of the parameters of eq. (7) requires that the dimension of the linear space $H^c_{SC,g}$ is at least 1.

Relative to the identification condition for the individual level model, there are some important differences. Specifically, in the aggregate case, one no longer has access to instrumental variables based on the averaging of individual-level variables. In order to achieve identification, it is necessary to have prior knowledge of aggregate variables that affect social capital but do not affect the aggregate outcome under study. Intuitively, in the aggregate data case, one is in essence working with a standard simultaneous equations system, so cross-equation exclusion restrictions must be employed to achieve identification.

To repeat, the import of these various econometrics issues depends on the context under study, the data available to a researcher, etc. The issues raised in this section should be regarded as providing benchmarks in the assessment of empirical studies; their salience will depend on the context that is under study.

IV.iii.c. Identification with predetermined social capital

When social capital is predetermined, the relevant individual level equation is now

$$\omega_i = k + cX_i + dY_{g(i)} + J_1 \mathrm{E}\left(\omega_{g(i)} \middle| F_{g(i)}\right) + J_2 SC_{g(i)} + \varepsilon_i \tag{9}$$

which means that social capital enters the equation in a symmetric way to the contextual effects $Y_{g(i)}$. Identification for models of this type has been initially studied in Manski (1993) and subsequently by Brock and Durlauf (2001a,b); an identification problem still exists because of the potential multicollinearity of $\mathrm{E}\left(\omega_{g(i)} \middle| F_{g(i)}\right)$ with the other control variables in (9). Durlauf (2002c) provides the following necessary conditions for identification.

Proposition 3. Identification of individual level behavioral equation with exogenous social capital

Identification of the parameters in eq. (9) requires

i. The dimension of the linear space spanned by elements of $\left(1, X_i, Y_{g(i)}, SC_{g(i)}\right)$ is $r+s+2$.

ii. The dimension of the linear space spanned by the elements of $\left(1, X_i, X_{g(i)}, Y_{g(i)}, SC_{g(i)}\right)$ is at least $r+s+3$.

However, unlike the endogenous social capital case, it may be possible to identify whether the role of social capital is nonzero even if (9) is not identified. Following an argument of Manski (1993), observe that the reduced form for (9) is

$$\omega_i = \frac{k}{1-J_1} + cX_i + \frac{J_1 c}{1-J_1} X_{g(i)} + \frac{d}{1-J_1} Y_{g(i)} + \frac{J_2}{1-J_1} SC_{g(i)} + \varepsilon_i \qquad (10)$$

Identification of the compound parameter $\frac{J_2}{1-J_1}$ is sufficient for determining whether there is some social capital effect. Identification of this parameter requires that the social capital variable is not linearly dependent on the other variables in (10); formally (Durlauf (2002c)) verifies

Proposition 4. Identification of a social capital effect when social capital is exogenous

If the dimension of $\left(1, X_i, X_{g(i)}, Y_{g(i)}, SC_{g(i)}\right)$ exceeds $\left(1, X_i, X_{g(i)}, Y_{g(i)}\right)$ then the presence of a social capital effect may be identified from (10).

Proposition 4 may be readily extended to the case of aggregate data; if aggregate social capital is exogenous then it is simply nothing more than an additional regressor in an aggregate outcome regression. On the other hand, if one is working with aggregate data and social capital is exogenous, then it is impossible to identify any of the model parameters. The reason is simple: there are no longer any instrumental variables available from the social capital equation to instrument $\mathrm{E}\left(\omega_{g(i)} \middle| F_{g(i)}\right)$, so no analog to Proposition 3 exists.

IV.iv. Additional issues

A number of difficulties beyond identification plague empirical work on social capital. As we have emphasized in Section II, reliance on interpersonal relationships and networks can often be seen as a symptom that formal institutions do not work well.[17] To

[17] This does not imply that networks would never be observed in well developed markets. Through interpersonal relationships, economic agents may form coalitions to subvert the market equilibrium to their advantage. Think of cartels, for instance. Clubs and networks can similarly be used to bias market outcomes, e.g., to ban non-whites or women from

illustrate how this might impact statistical analysis, suppose we have data on labor markets in different countries and we seek to estimate whether the density of social networks raises the average quality of the match between workers and employers. Suppose for the sake of argument that we have a convincing measure for the average quality of the match. Regressing this measure against the density of social networks is likely to yield incorrect results if the researcher does not control for differences in formal institutions across the countries.

For instance, employment offices may play an active match-making role in some countries. Failing to control for employment offices would underestimate the effect of social capital. In fact, if employment offices channel information more efficiently than interpersonal networks and if these networks arise in response to the absence of employment offices, countries with more networks will have less efficient labor markets.

Studies of the effects of social capital on the delivery of public goods suffer from other problems as well. Earlier in this section we have argued that social capital is difficult to disentangle from other group effects. One such group effect likely to influence empirical work is the role of leadership. Community leaders often play a crucial role in fostering the creation of social capital – e.g., membership drive – that they can harness for a particular goal. Observing a relationship between social capital and the presence of a public good may be due to the presence of a third, unobserved factor: leadership. The distinction between the two effects is important for policy because good community leaders are rare and leadership is much harder to replicate than groups.

V. Empirical studies of the effects of social capital

Following the econometric discussion, the literature on the effects of social capital may be divided into two types: individual and aggregate studies.

V.i. Individual-level studies

---

certain jobs. Political clientelism is another example (Bayart (1989)). In all these cases, social capital actually reduces aggregate welfare.

Individual-level studies of social capital may be divided into studies that focus on developing societies and studies that focus on OECD societies. This division reflects more than data sets. Studies of social capital in developing societies are associated with somewhat different questions than their OECD (primarily United States-based) counterparts. This division reflects differences in underlying concerns. Development scholars are interested in social capital as a mechanism to ameliorate society-wide problems whereas interest in advanced societies tends to derive from concerns about the persistence of social exclusion and poverty in affluent societies.

A typical social capital study in this literature posits an individual outcome of the form

$$\omega_i = \gamma X_i + \pi Y_{g(i)} + JSC_{g(i)} + \varepsilon_i \tag{11}$$

where, following previous notation, X_i denotes a set of individual controls, $Y_{g(i)}$ denotes a set of group controls and $SC_{g(i)}$ denotes social capital. As such, eq. (11) corresponds to the case of exogenous social capital discussed in Section III. Evidence for the relevance of social capital is equated with the statistical significance of the coefficient *J*. In the various tables we have constructed to summarize various empirical papers, we report dependent variables and social capital measures, as well as findings based on the statistical significance standard.

Social capital and development

Links between social capital and development have been examined in a range of contexts. One reason for this is that the failure of many developing economies to achieve sustained growth has led social scientists to look for previously unexplored factors in the development process. Table 1 lists a number of studies of social capital in developing societies.

As the table indicates, a range of alternative outcomes have been studied. Similarly, a range of social capital measures have been employed. While these studies are quite disparate, there are some commonalities. First, these development studies typically focus on measures describing the social networks in which individuals participate. Fafchamps and Lund (2003), Fafchamps and Minten (2001,2002), Grootaert (2000), Isham (2002) and Narayan and Pritchett (1999) all give primary focus to the role of memberships in various organization and trading networks as determinants of economic outcomes. The quite different social capital measures used by Lee and Brinton (1996) and Palloni *et al* (2001) reflect the different outcomes they are measuring (immigration and placement in elite firms.) Further, the studies in Table 1 give primary focus to participation in organizations that can provide economic benefits in terms of information sharing and the production of collective goods. In this sense, these studies focus on economic benefits to organizations as opposed to more tangible psychological and social benefits.

From the perspective of the discussion of identification in Section III, several questions arise. First, how does one differentiate social capital effects from the presence of other group effects such as information spillovers, or the presence of common factors such as legal or political institutions? In the papers discussed here, relatively little attention has been paid to this question. Notice that the failure to consider this issue is not necessarily a damning criticism, in the sense that one may have reasons to rule out such effects in advance. However, these studies also typically fail to make good arguments that alternative social determinants of outcomes can be ignored. This strikes us as a more serious indictment in that social capital variables can easily proxy for such factors. Put differently, we have argued that social capital represents a new explanation of individual and aggregate outcomes primarily to the extent that it embodies certain types of informal norms. The empirical literature typically does not contrast this view with alternative perspectives on social interactions.

In our judgment, the more successful studies of social capital and development are those that have focused on specific phenomena that have been placed under the social capital rubric. Unsurprisingly, Fafchamps and Minten (2002) is in our view a good example of this approach. As indicated in the paper's title, the focus of the analysis is

less on social capital per se than on the role of social networks in affecting trader profitability. This paper focuses on agricultural traders in Madagascar. These traders are intermediaries between farmers and various markets in the country. Because the goods they sell (staples such as rice, potatoes, and beans) are well defined (the basic goods are homogeneous and are distinguishable by observable features such as whether they have been milled or converted to flour, etc.), it is relatively easy to measure the value added associated with a trader's activity. Fafchamps and Minten (2002) find that measures of the size of an individual trader's business network are positively associated with value added and total sales. The paper argues that a relationship between networks and these economic outcomes may be understood in the context of models of imperfect information and monitoring, which provides a clear theoretical motivation for the empirical framework as well as a plausible theoretical interpretation for the various findings.

Finally, it should be noted that while the different studies in Table 1 consistently support a role for social capital in facilitating various economic outcomes, two of the studies, Krishna (2001) and Varughese and Ostrom (2001), argue that there are important subtleties in this relationship that need to be accounted for. Krishna (2001) finds that for villages in Rajastan India, the relationship between conventional social capital measures and outcomes such as common land development and poverty reduction is sensitive to a notion of effective governance Krishna calls "capable agency." By capable agency, Krishna refers to factors such as strong leadership in organizations, frequent interactions between villagers and clients, etc. His argument is that the density of organizations, a variable often used to measure social capital, will be associated with socially better outcomes only when capable agency is present. Varughese and Ostrom (2001) find, based on a study of groups of forest users in Nepal, that levels of collective action are not well predicted by measures of ethnic, caste, and religious homogeneity within these groups. These sorts of variables are often used to proxy for social capital. Varughese and Ostrom (2001) conclude that institutional design, how decisions are made, etc, can overcome barriers to cooperation that are induced by heterogeneity. Taken together, these studies illustrate that successful group activities depend on more than the presence of social ties per se.

Social capital in OECD societies

Just as social capital has been used to explain a range of outcomes in developing economies, so it has been used to explain a range of US phenomena. Table 2 reports a number of such studies.

In comparing Tables 1 and 2, a number of differences may be identified. First, social capital studies for affluent societies are far more heterogeneous than those which we report for developing economies. One finds studies of social capital for the United States that explore outcomes ranging from mental health (Furstenburg and Hughes (1995)) to dropping out of high school (Teachman, Paasch, and Carver (1997)) to criminal activity (Hagan and McCarthy (1995)). We do not believe this reflects differences in our choices of what studies to report. Rather, interest in social capital in advanced societies has been motivated by different phenomena than in the case of developing economies. In particular, the focus on social capital appears to be motivated by a desire to understand how some individuals avoid self-harming behaviors of various types.

Second, social capital studies for affluent societies focus on somewhat different variables to proxy for social capital than their development counterparts. This may be seen in the frequent examination of parental influences in Table 2. A common assumption in studies for the US is that the parent, child, neighborhood and school relationships are a primary form of social capital. McNeal (1999), for example, explicitly argues that parent/child interactions closely correspond to what Coleman originally meant by social capital.

Another feature that distinguishes the literature on OECD societies is its focus on traditionally sociological concepts in construing social capital. One important notion is intergenerational closure, which holds when parents of a given child know both his friends as well as his friends' parents; both Morgan and Sorenson (1999a) and Sandefur, Meier, and Hernandez (1999) treat closure as an important aspect of social capital. This variable arises because, as argued originally in Coleman (1988), control and monitoring of children is sensitive to the ways that a family is embedded in a community.

While OECD social capital studies typically are based on richer data sets than those available for developing countries, these studies often suffer from serious flaws. One problem is that little discipline has been imposed on the empirical proxies used for social capital, which makes many of the empirical claims in this literature incredible. For example, authors such as Furstenburg and Hughes (1995), McNeal (1999) and Sandefur, Meier, and Hernandez (1999) treat the number of family moves as a measure of social capital for youths. The idea is that the more a family moves, the weaker the social ties between the youth and his community. This is certainly a plausible claim. However, it does not suffice to make family moves a valid social capital measure. Since moves are endogenous, the variable in essence provides an indictor for those characteristics that determine the moves. Such characteristics can be associated with different youth outcomes for reasons that have nothing to do with social capital. For example, families who make more moves plausibly contain parents who are less interested in their children than those who make fewer, since such parents may be putting less weight on the costs to children of changing neighborhoods. Parents with less interest in their children (which can be formalized by using Loury's (1981) model of intergenerational mobility and allowing for heterogeneity in the rates at which parents discount offspring utility) will presumably invest less in their children, altering their outcomes in ways similar to the purported effects of lower social capital. Our point is not that one explanation or the other is correct, but rather that neither is identified from the data. Put differently, there are good reasons to believe that there are systematic differences in the unexplained components of individual behavior that render standard estimation methods inconsistent; specifically, families asserted to posses high levels of social capital, from the perspective of the estimated model, may be expected to be associated with higher levels of parental interest in children, which means the residuals in the associated regressions no longer have conditional expectations of 0. As such, this discussion is an illustration of an exchangeability violation of the type discussed in Section III; Furstenberg and Hughes (1995) are especially susceptible to this criticism due to the lack of attention to control variables.

Similarly, little attention is typically given to the identification problem of distinguishing social capital from endogenous or other group effects. This failure derives

from the flexibility of the social capital definitions that are employed. Is a psychological propensity to behave similarly to one's peers a form of social capital? The answer to this question is unclear from the literature, since such a propensity could easily count as a type of social norm.

While none of the studies in Table 2 can be said to fully address these general statistical questions, some of the studies are nevertheless clearly valuable contributions. One paper we would identify is Morgan and Sorenson (1999a). This paper is noteworthy for its careful attention to different causal mechanisms by which social capital may matter and by the care with which empirical proxies are constructed. We would also note that the paper focuses on a very specific issue, namely why Catholic schools appear to outperform their public counterparts, where there are good prior reasons to believe social factors matter.[18] Palloni *et al* (2001) is in many ways a very different study, yet is also very admirable. This analysis focuses on a very simple notion of social capital, in studying the effect on an individual's migration decision of prior migration by a sibling. What commends this study is the immense care taken to deal with questions of unobserved heterogeneity and common factors between siblings unrelated to social capital.

Before leaving this section, we draw attention to Costa and Kahn (2003b), which provides an historical perspective on social capital. In this paper, the behavior of union soldiers in the Civil War is examined, with particular attention to rates of promotion and desertion across different companies of soldiers. Costa and Kahn find that ethnic and occupational homogeneity of companies was conducive to braver conduct by soldiers. While far removed from the types of behaviors that are usually studied using social capital, the behavior of soldiers is in fact an excellent phenomenon to examine, given the

[18] Morgan and Sorenson (1999a) has in fact engendered some controversy, see Carbonaro (1999) and Hallinan and Kubitschek (1999). The main thrust of these criticisms concerns the extent to which the social closure measures used by Morgan and Sorenson fully capture the relevant social dynamics. We believe that the rejoinder Morgan and Sorenson (1999b) effectively answers these objections; equally important, these objections do not mitigate the reasons we admire the study. The level at which debate on this paper occurred is far deeper than the great majority of efforts to link social capital concepts to data.

well documented role of social factors in battlefield conduct.[19] We believe creative exploration of data sets like this can add a great deal to the understanding of social capital.

V.ii. Aggregate studies

At the beginning of Section III, we outlined the difficulty of estimating the beneficial effects of social capital from individual data. We now turn to empirical studies that rely on aggregate data and examine whether they provide more convincing evidence of social capital. Table 3 reports a number of social capital studies that employ such data. As the Table indicates, a large number of aggregate level social capital studies have focused on the relationship between social capital and per capita output growth at a high level of aggregation, such as a country or region. As such, most of the studies of this type are variants on empirical growth regressions that have become a workhorse of modern growth economics.[20] An assessment of the aggregate studies using social capital is therefore essentially equivalent to an assessment of a set of growth regressions designed to establish that a particular variable is causally related to growth.

Growth regressions of the type found in the studies of Table 3 have been subjected to very serious methodological criticisms; examples include Brock and Durlauf (2001b), Durlauf (2000), Durlauf and Quah (1999), and Temple (2000). As argued in these papers, growth regressions suffer from several fundamental problems that make implausible the types of causal inferences one typically finds in the empirical literature. First, there is the problem of the choice of control variables. Growth theories are open-ended, which means that one growth theory does not have any logical implications for the truth or falsity of another. Hence, there is no natural way, when one wishes to test the importance of a given theory, to identify the appropriate set of theories to incorporate in a correctly specified structural growth model. As Durlauf and Quah (1999) indicate, there

[19]To be clear, social factors can play a negative role in military behavior, such as in violence against civilians. See Aaronson (1999) for discussion of the social dynamics that occurred among US soldiers during the My Lai massacre of Vietnamese civilians.

[20]See Durlauf and Quah (1999) and Temple (1999) for surveys of the methods and findings of the empirical growth literature.

are in fact more extant growth theories than there are countries to which they are supposed to apply. As a result, any given growth regression may be subjected to the criticism that relevant control variables have been omitted. While there are some possible ways to deal with this problem, see Fernandez, Ley, and Steel (2001), this problem has not been addressed in any social capital and growth studies, as far as we know.

Second, growth regressions typically fail to account properly for parameter heterogeneity across countries. Evidence of such heterogeneity may be found in Desdoigts (1999), Durlauf and Johnson (1995), and Durlauf, Kourtellos, and Minkin (2001); theoretical models that imply heterogeneous growth processes for different groups of countries include Azariadis and Drazen (1990) and Howitt and Mayer-Foulkes (2002). Failure to account for parameter heterogeneity calls into question the structural interpretation of a social capital variable as it may be proxying for this form of heterogeneity. One example that is suggestive of this possibility concerns the role of ethnic heterogeneity in growth, a question studied by Easterly and Levine (1997).[21] In this paper, the authors argue that ethnic conflict inhibits public good creation and so acts as an impediment to growth. Ethnic conflict is instrumented with a measure of ethnolinguistic diversity which proves to be strongly negatively associated with growth. Since sub-Saharan Africa has exceptionally high levels of ethnolinguistic diversity, the authors conclude that this is an important mechanism in understanding Africa's growth problems. Brock and Durlauf (2001a) reexamine this study, allowing for various types of exchangeability violations due to parameter heterogeneity, and find that the relationship between ethnolinguistic diversity and growth appears only for sub-Saharan Africa; this variable does not help explain growth patterns in the rest of the world. Brock and Durlauf's finding illustrates how growth explanations may well not be constant across countries. And for the African case, it is unclear whether the growth findings are causal or whether ethnolinguistic diversity simply proxies for some other form of "African exceptionalism."

[21]It should be noted that Easterly and Levine (1997) does not explicitly focus on social capital; however, the mechanisms by which ethnic heterogeneity can affect economic performance are in many cases the same as have been proposed in the social capital literature.

Taken as a whole, these arguments imply that the social capital/growth studies do not meet the exchangeability requirements that we discussed in Section III. While this reflects more general failings of the empirical growth literature (Brock and Durlauf (2001b)), it is also the case that growth studies using social capital have been quite insensitive to efforts in the growth literature to address these problems.

Beyond questions concerning the comparability of observations, there are unresolved issues concerning causal interpretation of growth regressions that apply to the social capital case. This is especially important given the endogeneity of aggregate measures of social capital. We are unaware of any social capital study using aggregate data that addresses causality versus correlation for social capital and growth in a persuasive way. While this is a broad brush with which to tar this empirical literature, we believe it is valid. A related problem is that we are unaware of any compelling instrumental variables for social capital in these regressions. This failure is a corollary of the absence of any strong theories of aggregate social capital determination in the social science literature that would allow one to characterize appropriate instruments.

When one turns from national-level growth studies to other aggregate studies, the plausibility of claims concerning social capital becomes stronger in some cases. A recent study by Goldin and Katz (1999) is particularly interesting in its focus on the sources for the rise of high school attendance in Iowa in the early part of the twentieth century. By focusing on characteristics of Iowa counties, they are able to avoid some of the clear problems of exchangeability that plague studies using coarser levels of aggregation. But even here, other problems arise: more important, the data available are quite weak in the sense that the variables which suggest the presence of social capital effects could equally well suggest alternative explanations. The specific variables that seem most suggestive of social capital effects are the percentage of native born citizens and the population of towns; high percentages of native born and low population sizes are each associated with higher high school attendance. Clearly, linking these correlations to a causal role for social capital or other type of social influence is speculative. To be fair, Goldin and Katz

(1999) point out that there may be alternative explanations, such as the smaller towns having fewer opportunities for those without high school educations.[22]

Overall, we conclude that aggregate social capital studies have not been successful in providing compelling empirical evidence on the effects of social capital. These studies require identifying assumptions that are incredible by conventional social science reasoning. We believe that research efforts should be directed towards micro-level studies as the problems with country-wide studies seem too intractable to overcome. Data at lower levels of aggregation, such as county data for a homogeneous place like 1915 Iowa, are likely to be more amenable to persuasive analysis, provided the issues of exchangeability and identification can be addressed adequately.

VI. Empirical studies of the level and determinants of social capital

Interest in the effects of social capital has spawned a related literature of the level of social capital and how this level is determined. Table 4 lists a range of studies that have explored this issue. It is worth noting that while attention has been given to questions of model specification and identification for models in which social capital is a causal determinant of various outcomes, we are unaware of any formal analyses that have been applied to models of social capital formation. Our conjecture is that the arguments applied to models of social capital effects can be extended in a straightforward fashion to models of social capital determinants, but this remains to be done.

One important question in the literature on the formation of social capital has been whether the extremely prominent claims by Putnam (1995,2000) that social capital in the United States has experienced a major decline are correct, and if so, whether this decline can be attributed to those factors he has described, namely, increased watching of television and the passing of the World War II generation. It appears that many of Putnam's claims have not withstood careful scrutiny. Paxton (1999) shows that there is

[22]At the other extreme, the effort by Robison and Siles (1999) to link aspects of state level income distributions to various social capital proxies fails to make any serious effort to ensure exchangeability; in addition the variables used to measure social capital, such as labor force participation, render the claims made about social capital untenable.

little evidence of secular declines of trust or overall associational activity in the United States. Bianchi and Robinson (1997) find little evidence that patterns of television viewing have much relationship to maternal employment status or other family factors often asserted to lead to lower social capital. Costa and Kahn (2003a), using more disaggregated measures of associational activity, find declines in social capital measures that are qualitatively similar to what Putnam has claimed. However, they find rather different explanations. Their analysis concludes that the decline in social capital produced "outside the home" such as volunteering is explained to a large extent by the rise in female labor force participation in the last 4 decades. This study also finds that declines in social capital produced "inside the home" such as frequency of socializing is strongly related to increases in neighborhood heterogeneity. One important implication of this work is that it places claims about a decline in US social capital in a different normative light. If increasing female labor force participation is due to the breakdown of discriminatory barriers against women in labor markets and if increasing neighborhood heterogeneity reflects a breakdown of the levels of social and ethnic segregation in the United States, then perhaps declines in social capital are best thought of as an unfortunate but necessary side effect of a movement towards a more just society and so should not be mourned.

One important aspect of this research is the move towards a causal understanding of the processes by which social capital is formed. One interesting example of such work is Brehm and Rahn (1997) who employ General Social Survey data to study the reciprocal interaction of community involvement and trust in others. Their analysis finds a stronger causal relationship between community participation to trust than the converse. This finding is indicative of the empirical importance of Dasgupta's (2002) argument that social capital should be modeled as a network.

Other studies have focused on identifying predictors of trust. For the US, Alesina and La Ferrara (2002) find that trust in others is negatively associated with community heterogeneity. Rahn and Rudolph (2002) extend work of this type in an analysis of the determinants of trust in local government. This paper finds that political culture and community heterogeneity play an important role in explaining trust. Interestingly, trust does not appear to be influenced by the form of local government as trust levels are not

predicted by whether a community has a mayor or city manager (the latter implying less popular control of local government). These studies are best regarded as reduced form analyses in that issues of causality are not specifically addressed.

An especially important effort to understand the formation of social capital is the Project on Human Development in Chicago Neighborhoods (PHDCN). This is a remarkably detailed data collection project that covers several hundred neighborhoods in Chicago. These data are proving to be very useful in delineating the detailed social structure of neighborhoods. As described in Sampson, Morenoff, and Earls (1999 pg. 639), the available data include responses to questions such as "About how often do you and people in your neighborhood do favors for each other?" and the likelihood that one's neighbors would intervene if one's child were observed skipping school.

Sampson, Morenoff, and Earls (1999) use the PHDCN to study a range of social aspects of neighborhoods. In particular, they distinguish the social capital of a neighborhood as "the resource potential of personal and organizational networks" (pg. 635) from the collective efficacy of a neighborhood, "a task-specific construct that relates to the shared expectations and mutual engagement by adults in the active support and social control of children." (pg. 635). The purpose of this distinction is to differentiate general notions of neighborhood social resources from the use of these resources. By delineating how neighborhood members help one another, for example through monitoring one another's children, Sampson, Morenoff, and Earls (1999) give a rich portrait of how neighborhoods benefit their members, illustrating how help in childrearing or trust among neighbors are important mediating variables in understanding why poor neighborhoods have adverse effects on their members. By uncovering specific mechanisms by which neighborhoods matter, this study moves beyond the common use of social capital variables in which the link between the variable and a behavioral outcome is metaphorical and all too often a black box.

VII. Suggestions for future research

As our discussion suggests, we believe that social capital studies have very often been unpersuasive. We make the following suggestions as to how one can improve this literature.

First, empirical analyses need to step back from grandiose approaches to social capital and focus on the more mundane but potentially far more fruitful task of analyzing specific social components to individual behavior. This does not require abandonment of social capital as a general organizing idea or metaphor, but rather means that evidence in favor of social capital should be derived from specific claims about social influences on individuals.

A useful contrast may be made between the Helliwell and Putnam (2000) paper, the study of regional differences in growth rates in Italy that we have criticized earlier, and a recent study by Glaeser, Laibson, Scheinkman and Soutter (2000) that explores the determinants of trust. Rather than run regressions that make incredible assumptions about the exchangeability of regional growth rates, Glaeser Laibson, Scheinkman and Soutter employ well crafted experiments to see how attitudes and background characteristics influence the choice of strategies in various economic experiments. In the context of these experiments, notions such as trust are quite well defined since it amounts to expectations about the play of other agents in the game. This well defined environment provides much more compelling evidence of how trust influences behavior than can be obtained from ad hoc regressions. The use of experiments to understand social capital is further developed in Carter and Castillo (2003,2004), who consider how variation in roles by players in economic experiments can allow for differentiation between altruism and trust as determinants of behaviors.

The importance of experimental evidence should not be exaggerated. Economic experiments are not a panacea for the limits of inference with observational data. One problem is generalizability; it is far from clear how behavior in economic experiments maps into behavior in the larger economy and society, although Glaeser Laibson, Scheinkman and Soutter make an important advance in this regard by attempting to correlate behavior in experiments with behavior in the "real world" by participants. Further, as discussed by Manski (2002) in an important recent paper, there are identification problems in experiments as it is often difficult to distinguish behavior that

is driven by altruistic preferences from behavior driven by selfish preferences but with expectations of trustworthy behavior by others. Nevertheless, Glaeser Laibson, Scheinkman and Soutter and Carter and Castillo represent a style of research that is an important advance in the social capital literature.

In addition, moving the discussion of social capital away from generalities to specific mechanisms in the way we suggest will allow one to deal with issues of endogeneity and exchangeability more effectively, since it will facilitate more precise and comprehensive modeling of causal mechanisms than one finds in the social capital literature. While the great majority of social capital studies include numerous control variables, the choice of these variables is rarely determined by careful delineation of the determinants of behavior of the agents under study. In addition, there has been little attention to questions of parameter heterogeneity.

A concrete implication of this discussion is that future research on social capital by the World Bank, for example, should be careful about the use of highly aggregated data. It is difficult to make compelling exchangeability arguments for data sets in which the observations are countries or regions. Ad hoc assumptions concerning the legitimacy of instrumental variables have plagued this literature for good reason: theories of social capital formation are underdeveloped so that it is difficult for researchers to sensibly construct aggregate measures of social capital.

Second, we believe that future data collection exercises must explicitly attempt to gather information on group-level influences, rather than on social capital alone. This should include measures of the quality of leadership. At the core of virtually all microeconomic reasoning is the general idea that decisions are purposeful outcomes based on an individual's preferences over outcomes, constraints on what actions are feasible, and beliefs over the consequences of those actions. The new social economics (cf. Durlauf and Young (2001)), is based upon the recognition that these three components to decisions are deeply influenced by social factors. A data collection exercise designed to explain a given set of outcomes should therefore be based on the

development of a typology of what sorts of social factors affect each of the components and the development of plausible empirical analogs to these social factors.[23]

The sorts of detailed data collection we advocate are in fact underway in some cases. In particular, the Project on Human Development in Chicago Neighborhoods and data collection based on the World Bank Social Capital Assessment Tool are exemplary. In each case, the levels of specificity in terms of uncovering how individuals interact in villages, communities and social networks is a great advance over the crude measures often used in social capital studies. The most obvious suggestion in terms of the design of these studies would be the exploration of the extent to which the existing survey questions are adequate in terms of dealing with the specification and identification problems we discuss in Section III. There is no quick answer to this as it would require integrating some theoretical modeling with the survey design. Nevertheless, the payoffs to such an endeavor could be quite high.

How does our admittedly very general advice differ from the way in which data collection on social capital is typically done? We have already discussed one difference, namely, the effectiveness of data collection is augmented when attention is paid to the uses to which the data will be applied. To repeat, the analysis of potential identification problems should inform data collection and not just define limits to which a data set may be used. Another important difference is that this approach avoids privileging social factors that can be construed as "social capital" over others. As we have argued, the failure to consider alternative social explanations to social capital is an important source of skepticism with respect to existing studies. More importantly, there is no *a priori* reason to assume that social capital is a more likely source of important effects than other social factors. Another difference is that our proposed approach, by separating social factors as concepts from empirical measurement, will avoid conflating the two, as often

[23] Sandefur and Laumann (1998) argue in favor of understanding social capital in terms of its benefits, identifying these as provision of information, influence and control in dealing with others, social solidarity between individuals. These types of benefits represent combinations of the preferences, constraints, and beliefs we advocate employing. An advantage of our approach is that our categories represent empirically meaningful differences in the determinants of individual behavior whereas the Sandefur and Laumann categories are necessarily interdependent and do not correspond to any

occurs. Finally, the exercise of modeling individual choice in order to determine what is meant by social factors should provide some guidance as to the appropriate levels at which these factors should be measured. Does an individual's or a society's level of trust matter for individual conduct? The appropriate answer to a question like this should derive from the decision problem at hand. Empirical studies of social capital have largely not addressed this question.

Third, there needs to be greater recognition of the limits to statistical analysis in contexts such as the evaluation of social capital. This is partly a restatement of the first suggestion in that there simply do not exist any available data or methodology that can allow an assessment of the broad claims of the sort one finds in the social capital literature. But beyond this, we believe economists need to be more receptive to the sorts of evidence found in other disciplines beyond the quantitative analyses that are standard in economics. For example, sustained descriptive histories can teach us much about the ways that social structures influence individual conduct even if they are not constructed in the form of claims about *F*-statistics and the like. At the other extreme, there is a wealth of information in the social psychology literature that addresses in precise ways the inchoate ideas about individual behavior that underlie the social capital literature. This suggestion requires greater openmindedness on the part of economists to nonstatistical sources of information. But the payoffs can be high both in terms of substantive understanding as well as in facilitating quantitative analyses. As the discussion of identification argued, social capital effects can only be revealed if one has prior information on what group effects do not directly influence individuals. This is information that nonstatistical studies may be able to provide.[24]

In fact, it is reasonable to argue that some aspects of the question of how social capital has facilitated socioeconomic or political development should be treated in the same spirit as questions such as what led to rise of emergence of democracy in ancient Athens versus a martial culture in ancient Sparta or what were the causes of World War I.

"natural kinds" in terms of either individual activity or collective action, at least as far as we can tell. For example, trust will affect information transmission.

[24] Of course, qualitative studies are not immune to the overinterpretation (due to ignoring identification problems) and overclaiming (due to exaggeration of the import of statistical

These are not meaningless questions; but it is necessary to accept limits as to the quantitative precision with which such questions can be answered and what it means to say the question has been answered. None of this suggests that statistical analysis should play anything other than a primary role in social capital studies; our argument is that the credibility of the social capital literature will be augmented when nonstatistical evidence is better used to motivate assumptions and suggest appropriate ways for formulating hypotheses.

VIII. Conclusions

In this Chapter, we have tried to provide an overview of the state of social capital research by both describing the state of the conceptual, theoretical and econometric literatures on social capital and by surveying a number of empirical studies. Our overall assessment of the social capital research is quite mixed. In terms of conceptual and theoretical studies of social capital, there is a considerable amount of ambiguity and confusion as to what social capital means. One conclusion we draw from our survey is that the most successful theoretical work on social capital is that which, following Dasgupta (2002), models social capital as a form of social network structure and uses the presence of that structure to understand how individual outcomes are affected in equilibrium. From the empirical perspective, the role of networks in facilitating exchange is one of the most compelling empirical findings in the social capital literature (cf. Fafchamps (2004)), so a more narrow focus on this type will likely not diminish the importance of social capital as a concept.

With respect to empirical work in general, social capital research has led to the development of a number of interesting data sets as well as the development of a number of provocative hypotheses, much of the empirical literature is at best suggestive and at worst easy to discount. So while one can point to no end of studies in which a variable that is asserted to proxy for social capital has some effect on individuals or groups, it is

---

findings taken on their own terms) that we have criticized in quantitative studies. See Tarrow (1996) for criticisms along these lines.

usually very difficult to treat the finding as establishing a causal role for social capital. We have highlighted a number of studies that we think are particularly strong, but those studies we find persuasive are relatively exceptional. The defects of the empirical social capital literature are unfortunate, since the work on social capital is an active front in which the "undersocialized conception of man" for which economics has been criticized (Granovetter (1985)) is being addressed.

One recommendation we make in regard to empirical studies is that social capital literature pay far more attention to formal issues of identification, self-selection and unobserved group characteristics. These issues have been extensively studied in the closely related context of social interactions (cf. Brock and Durlauf (2001c)) and many ideas from that literature may be applied to social capital. In addition, we believe that empirical social capital studies must do a much better job of differentiating between social capital effects and alternative types of group effects.

Attempts to provide social richness to economic analysis will only succeed if the theoretical and empirical work that accompanies this effort is subjected to the same rigorous standards that are required of other analyses in economics. In contrast, the extravagant claims so often found in this literature (an outstanding example of which is Putnam (2000) who appears capable of attributing every conceivable societal virtue to social capital)[25] have little prospect of having lasting social science value. Beyond the failure to contribute to the social science enterprise, there is a legitimate concern that studies which make excessive claims and unsupported assertions can discredit social capital as an idea. In conclusion, what the empirical social capital literature ultimately needs is more matter and less art.

---

[25]See Durlauf (2002b) for an extended critique of Putnam (2000) which addresses the problem of overclaiming, faulting Putnam both for not dealing with some of the identification problems we hve described in Section IV as well as for failing to analyze social capital in a fashion conducive to rigorous policy analysis.

TABLE 1: INDIVIDUAL-LEVEL STUDIES OF SOCIAL CAPITAL IN DEVELOPING COUNTRIES

STUDY	AGENTS	OUTCOMES	SOCIAL CAPITAL MEASURES	FINDINGS
Carter and Maluccio (2003)	Households in KwaZulu-Natal South Africa	Child height for age	Number of associations in community and interaction of family income with community income	Social capital helps ameliorate effects of individual-specific economic shocks
Fafchamps and Minten (2002)	Food traders in Madagascar	Value added and total sales	Number of traders known, number of relatives in agricultural trade, number of potential informal traders	Number of traders known and number of potential informal traders statistically significant.
Grootaert (2000)	Rural households in Indonesia	Per capita household expenditure	Number of memberships in associations, diversity of memberships, number of meetings of associations, index of participation in decisionmaking, measure of cash contribution to associations, measure of time contribution to association, measure of orientation towards community.	Social capital index statistically significant; number of memberships, internal heterogeneity of associations and level of participation in decisionmaking appear most important.

TABLE 1: INDIVIDUAL-LEVEL STUDIES OF SOCIAL CAPITAL IN DEVELOPING COUNTRIES

STUDY	AGENTS	OUTCOMES	SOCIAL CAPITAL MEASURES	FINDINGS
Isham (2002)	Households in rural Tanzania	Adoption of improved fertilizer	Village level measures of ethnic homogeneity for organizations in which households are members, levels of participation of household in organization decisionmaking, and extent to which leaders of village organization have different livelihoods than village members	Social capital measures are generally statistically significant predictors of adoption, but some regional differences exist
Krishna (2001)	Villages in Rajastan, India	Performance with respect to common land development, poverty reduction, and employment	Survey measures of participation in labor-sharing groups, trust, solidarity, and reciprocity	Efficacy of social capital is related to strength of leaders of associations, patron-client relations, etc.
Krishna and Uphoff (1999)	Villages in Rajastan, India	Collective action to restore degraded or vulnerable common lands	Social capital index based on survey answers to questions on level of collective action in village, village governance, village sense of obligation, etc.	Index is a strong predictor of better development outcomes

TABLE 1: INDIVIDUAL-LEVEL STUDIES OF SOCIAL CAPITAL IN DEVELOPING COUNTRIES

STUDY	AGENTS	OUTCOMES	SOCIAL CAPITAL MEASURES	FINDINGS
Lee and Brinton (1996)	Graduates of elite colleges in South Korea	Employment opportunities at large firms	Private social capital (family and friendship ties) and institutional social capital (social ties provided by university, eg. introductions to firms)	Institutional rather than private social capital is important in determining employment opportunities
Malucccio, Haddad, and May (2001)	Households in Kwazulu-Natal Province, South Africa	Per capita total expenditure	Index of individual memberships in groups, reflecting number, gender heterogeneity, and performance, based on survey responses. Community social capital levels computed as aggregates of individual indices	Individual and community social capital measures statistically significantly associated with expenditure in 1998 but not 1993
Narayan and Pritchett (1999)	Households in rural Tanzania	Per capita household expenditure	Social capital indices constructed for both households and villages. Indices based on memberships in groups, characteristics of the groups, and household values and attitudes	Village social capital dominates individual social capital
Palloni, Massey, *et al* (2001)	Sibling pairs in Mexico	Migration to the United States	Previous migration of one sibling	Likelihood of migration is increased if a sibling has already migrated

TABLE 1: INDIVIDUAL-LEVEL STUDIES OF SOCIAL CAPITAL IN DEVELOPING COUNTRIES

STUDY	AGENTS	OUTCOMES	SOCIAL CAPITAL MEASURES	FINDINGS
Pargal, Huq, and Gilligan (1999)	Households in Dhaka, Bangladesh	Establishment of voluntary solid waste management (VWSM) systems for neighborhoods	Indices of trust, reciprocity, and sharing for neighborhoods	Reciprocity index is best predictor of likelihood that a neighborhood has VWSM system
Varughese and Ostrom (2001)	Groups of forest users in Nepal	Level of collective activity, monitoring of forest use, enforcement of harvesting constraints, etc.	Homogeneity within group in wealth, caste, ethnicity	No necessary relationship between homogeneity and level of collective action; institutional design is more important

TABLE 2: INDIVIDUAL-LEVEL STUDIES OF SOCIAL CAPITAL: OECD COUNTRIES

STUDY	ACTORS	OUTCOMES	SOCIAL CAPITAL MEASURES	FINDINGS
Costa and Kahn (2003b)	Union soldiers in the US Civil War	Performance over course of war in terms of promotions, desertion, etc.	Homogeneity of companies of soldiers with respect to ethnicity, occupation, and age	More homogeneous companies are associated with more promotions and lower rates of desertion
Fernandez, Castilla, and Moore (2000)	Phone center employers	Returns to investments	Use of employees social networks in making new hires	Investment in use of employee referrals is shown to be quite profitable
Frank and Yasumoto (1996)	French financial elite; i.e. prominent individuals associated with financial institutions	Business dealings with one another	Reciprocity, trust. Actors are organized into subgroups based on friendship ties. Trust, equated with absence of hostile business actions, such as a hostile takeover, is expected to be higher between members of common subgroup. Reciprocity, defined as supportive actions such as helping a firm fend off a hostile takeover is expected to be higher between subgroups.	Basic predictions confirmed.

TABLE 2: INDIVIDUAL-LEVEL STUDIES OF SOCIAL CAPITAL: OECD COUNTRIES

STUDY	ACTORS	OUTCOMES	SOCIAL CAPITAL MEASURES	FINDINGS
Furstenberg and Hughes (1995)	Children of teenage mothers (US)	Graduation from high school, college enrollment, economic status, avoidance of live birth, avoidance of criminal activity, mental health	Within family social capital (presence of father in home, parents' expectations for school performance, etc.), family links to community (religious involvement, help network, neighborhood quality, etc.)	Various outcomes and social capital measures statistically significantly associated, even controlling for some human capital measures
Guiso, Sapienza, and Zingales (2002)	Households in Italy	Financial activities such as use of formal credit, portfolio behavior	Electoral participation and blood donation and province level	Social capital measures for both current location and place of birth predict use of formal credit, and investment in stocks rather than cash. Effects stronger for the poorer and less educated.
Hagan, MacMillan, and Wheaton (1996)	Teenagers in Toronto	Level of educational attainment, occupational status	Parental involvement with children, family moves across neighborhoods	Both types of social capital statistically significant in predicting outcomes
Hagan and McCarthy (1995)	Teenagers (Canada)	Various forms of criminal behavior	Social variables such as criminal mentors and criminal social networks	Social variables predict criminality

TABLE 2: INDIVIDUAL-LEVEL STUDIES OF SOCIAL CAPITAL: OECD COUNTRIES

STUDY	ACTORS	OUTCOMES	SOCIAL CAPITAL MEASURES	FINDINGS
McNeal (1999)	Teenagers in US	Academic achievement in science, truancy, staying in school	Parental interactions with child and with school	Favorable social capital effects on child outcomes seem only to apply to white students from middle and upper class backgrounds
Morgan and Sorenson (1999a)	Teenagers in US	Test scores in mathematics	Social closure around school, parental involvement in school, parental knowledge of friends	Social closure is negatively associated with test scores, in contradiction to standard predictions of social capital analyses
Parcel and Menaghan (1993)	Children in US	Index of child behavioral problems	Miscellaneous measures of family structure, parents' working conditions, and parents' personal resources, such as sense of self-estimation	Role of family social capital generally confirmed through statistical significance
Sandefur, Meier, and Hernandez (1999)	Teenagers in US	Intergenerational closure, parent/child interactions, high school graduation, post-secondary enrollment, enrolling in a four-year college	Family structure, number of times child changed schools, Catholic High school attendance	Various social capital measures are associated with outcomes in ways predicted by theory.

TABLE 2: INDIVIDUAL-LEVEL STUDIES OF SOCIAL CAPITAL: OECD COUNTRIES

STUDY	ACTORS	OUTCOMES	SOCIAL CAPITAL MEASURES	FINDINGS
Sun (1999)	Teenagers (US)	Academic performance measured by test scores	Structural measures (number of school changes, family structure) and process variables (parent child interactions, participation in activities, number of parents known),	Various process variables associated with test scores.
Teachman, Paasch, and Carver (1997)	Teenagers (US)	Dropping out of high school	Family social capital (living arrangements with parents, intensity of interactions with parents), community social capital (attendance in Catholic school, number of changes in school, measures of interactions of parents with schools and friends)	Attending a Catholic school and family structure robustly statistically significant across alternative specifications

TABLE 3: AGGREGATE-LEVEL STUDIES OF SOCIAL CAPITAL

STUDY	UNITS	OUTCOMES	SOCIAL CAPITAL MEASURES	FINDINGS
Beugelsdijk and van Schalk (2001)	European Regions	Per capita output growth	Trust, group participation	Group participation helps explain growth, but not trust
Easterly and Levine (1997)	Nations	Per capita output growth	Ethnic heterogeneity measured by ethnolinguistic diversity within a country	Per capita growth negatively associated with ethnolinguistic heterogeneity; important in explaining poor performance of sub-Saharan Africa
Goldin and Katz (1999)	Iowa Counties in 1915	High school attendance	Population size of towns, density of religious organizations, percentage of population that is native born	Small towns led expansion of high school attendance. Positive relationship with other possible social capital variables
Helliwell (1996)	Asian nations	Per capita output growth	Participation in associations, trust	Social capital measures contribute little once other factors such as openness are accounted for

TABLE 3: AGGREGATE-LEVEL STUDIES OF SOCIAL CAPITAL

STUDY	UNITS	OUTCOMES	SOCIAL CAPITAL MEASURES	FINDINGS
Helliwell and Putnam (2000)	Regions in Italy	Per capita output growth	Measure of civic community (index of associations, newspaper readership, and political behavior), institutional performance, citizen satisfaction with government	For the various measures, higher social capital associated with higher growth
Knack and Keefer (1997)	Nations	Per capita output growth	Indices of civic cooperation (measuring questions such as whether it is ever justified to cheat on taxes) and trust (percentage of individuals who say most people can be trusted)	Social capital measures help predict growth
LaPorta *et al* (1997)	Nations	Government efficiency (level of corruption, etc.), participation in politics and associations, social efficiency (infrastructure quality, infant mortality, educational level, etc.)	Trust	Trust generally statistically significant

TABLE 3: AGGREGATE-LEVEL STUDIES OF SOCIAL CAPITAL

STUDY	UNITS	OUTCOMES	SOCIAL CAPITAL MEASURES	FINDINGS
Lochner, Kawachi, Brennan, and Buka (2003)	Chicago neighborhoods	Aggregate and disease-specific mortality rates for neighborhoods and gender and ethnic groups within neighborhoods	Measures of trust, reciprocity, group participation	Social capital measures help to predict white mortality; relationship with mortality of blacks is weaker
Paxton (2002)	Nations	Index of liberal democracy	Number and types of international nongovernment organization in country, trust	Democracy and social capital reciprocally related; number of trade unions, sport associations and religious organizations negatively associated with democracy, number of others positively associated
Robison and Siles (1999)	US states	Means and coefficients of variation for household income	Measures of family structure, educational achievement, crime and labor force participation	Higher social capital proxies generally associated with higher means and lower dispersion in household income
Zak and Knack (2001)	Nations	Per capita output growth	Trust	Trust predicts growth even when factors such as property rights are controlled for.

TABLE 4: STUDIES OF SOCIAL CAPITAL FORMATION AND THE LEVEL OF SOCIAL CAPITAL

STUDY	AGENTS	SOCIAL CAPITAL MEASURES	POTENTIAL DETERMINANTS	FINDINGS
Alesina and La Ferrara (2002)	Adults in US	Trust	Miscellaneous personal and community characteristics	Low social capital measures for individuals are associated with membership in groups that have experienced discrimination (e.g. being African American), lack of economic success, community heterogeneity, experience of personal trauma
Bianchi and Robinson (1997)	Pre-teenagers in California	Time spent on studying and activities other than watching television	Family structure, parental characteristics, mother's labor force status	Study is higher and television watching lower among children of better educated; children of working mothers watch less television than others

TABLE 4: STUDIES OF SOCIAL CAPITAL FORMATION AND THE LEVEL OF SOCIAL CAPITAL

STUDY	AGENTS	SOCIAL CAPITAL MEASURES	POTENTIAL DETERMINANTS	FINDINGS
Brehm and Rahn (1997)	Adults in US	Civic engagement and civic trust	Reciprocal relationship between engagement and trust, confidence in institutions, life satisfaction, ethnicity, socioeconomic status, and many others	Participation strongly affects trust, each positively associated with socioeconomic status, confidence, negatively associated with being black
Charles and Kline (2002)	Adults in US	Carpooling	Ethnicity of neighbors	Ethnic heterogeneity reduces social capital formation for some pairings, notably whites and blacks and whites and Hispanics
Costa and Kahn (2003a)	Adults in US	Volunteering, socializing, non-church memberships,	Gender, community characteristics (race and income heterogeneity)	Declines in social capital produced outside the home such as volunteering are strongly related to higher female labor force participation; declining social capital within home such as frequency of socializing is strongly related to higher community heterogeneity

TABLE 4: STUDIES OF SOCIAL CAPITAL FORMATION AND THE LEVEL OF SOCIAL CAPITAL

STUDY	AGENTS	SOCIAL CAPITAL MEASURES	POTENTIAL DETERMINANTS	FINDINGS
DiPasquale and Glaeser (1999)	Adults in US	Citizenship (voting in local elections, helping solve local problems, knows school head, etc.)	Home ownership	Homeownership helps predict a range of citizenship variables.
Fafchamps (2003)	Traders in Benin, Madagascar, and Malawi	Trust in trading relationships	Ethnicity and religious similarity, gender, network effects	Ethnicity, religion and gender appear to have little effect on trust. Individuals possessing large numbers of business contacts give and receive more trust.
Gugerty and Kremer (2002)	Women's groups and school development projects in western Kenya	For women's groups, group size, attendance, financial status, and level of interactions with other groups and individuals; For schools, participation in school development projects	Funding of groups and funding of school textbooks.	Grants to women's groups appear to have had little effect on the capacities or size of women's groups; grants to governing committees of schools and increases in textbook funding were associated with increased participation of parents in school development; additional effects were found for textbook funding

TABLE 4: STUDIES OF SOCIAL CAPITAL FORMATION AND THE LEVEL OF SOCIAL CAPITAL

STUDY	AGENTS	SOCIAL CAPITAL MEASURES	POTENTIAL DETERMINANTS	FINDINGS
Hofferth, Boisjoly, and Duncan (1999)	Adults in US	Access to time and financial assistance from relatives and friends	Previous provision of time and financial assistance to those same relatives and friends	Time and assistance from friends is predicted by past provision, but not time and assistance by relatives
Miguel, Gertler, and Levine (2001)	Districts in Indonesia	Density of community organizations	Rapid industrialization within district	Industrialization, if anything was associated with rising density of organizations. Districts that neighbored districts experiencing rapid industrialization exhibited some declines, possibly due to out-migration
Oliver (1999)	Adults in US	Local civic participation	Community affluence and associated levels of social needs, competition for resources induced by population heterogeneity	Heterogeneous, middle income cities exhibit higher levels of civic participation than heterogeneous, affluent cities
Paxton (1999)	Adults in US	Trust, participation in various associations	Time	No strong evidence of declines in social capital in the US since the 1970's

TABLE 4: STUDIES OF SOCIAL CAPITAL FORMATION AND THE LEVEL OF SOCIAL CAPITAL

STUDY	AGENTS	SOCIAL CAPITAL MEASURES	POTENTIAL DETERMINANTS	FINDINGS
Rahn and Rudolph (2002)	Adults in US	Trust in local government	Measures of political institutions, political culture, income inequality, ethnic fractionalization, ideological polarization, controls for individual characteristics	Ideological polarization, income inequality, and political culture are more important and political institutions in explaining variation in trust
Sampson, Morenoff, and Earls (1999)	Adults in Chicago	Intergenerational closure, reciprocal social exchange, and shared expectations for informal social control	Miscellaneous neighborhood characteristics	Residential stability and relative affluence predict intergenerational social closure and reciprocal exchange, whereas neighborhood disadvantage predicts low expectations of shared child control

References

Aaronson, E., (1999), *The Social Animal, Eighth Edition*, New York: Worth Publishers.

Akerlof, G.A. and R. E. Kranton, (2000) "Economics and Identity", *Quarterly Journal of Economics*, 115, 3, 715-753.

Alesina, A. and E. La Ferrara, (2002), "Who Trusts Others?," *Journal of Public Economics*, 85, 207-234.

Annen, K., (2003), "Social Capital, Inclusive Networks, and Economic Performance," *Journal of Economic Behavior and Organization*, 50, 449-463.

Aristotle, (1985), *Nicomachean Ethics*, T. Irwin trans., Indianapolis: Hackett Publishing..

Arrow, K., (2000), "Observations on Social Capital," in *Social Capital: A Multifaceted Perspective*, P. Dasgupta and I. Seragilden, eds., Washington DC: World Bank.

Azariadis, C. and A. Drazen, (1990), "Threshold Externalities in Economic Development," *Quarterly Journal of Economics*, 105, 501-526.

Barr, A. (2000) "Social Capital and Technical Information Flows in the Ghanaian Manufacturing Sector," *Oxford Economic Papers*, 52, 3, 539-59.

Barr, A., M. Fafchamps, and T. Owens, (2004), "The Resources and Governance of Non-Governmental Organizations in Uganda", CSAE Working Paper, Oxford University

Bayart, J-F, (1989), *L'Etat en Afrique: La Politique du Ventre*, Paris: Fayard.

Bernstein, L. (1992), "Opting Out of the Legal System: Extralegal Contractual Relations in the Diamond Industry," *Journal of Legal Studies*, 21, 115-157

Bernstein, L. (1996), "Merchant Law in a Merchant Court: Rethinking the Code's Search for Immanent Business Norms," *University of Pennsylvania Law Review*, 144, 5, 1765-1821.

Beugelsdijk, S. and T. van Schalk, (2001), "Social Capital and Regional Economic Growth," mimeo, Tilburg University.

Bianchi, S. and J. Robinson, (1997), "What Did You Do Today? Children's Use of Time, Family Composition, and the Acquisition of Social Capital," *Journal of Marriage and the Family*, 59, 332-344.

Bigsten, A., P. Collier, S. Dercon, M. Fafchamps, B. Gauthier, J.W. Gunning, A. Isaksson, A. Oduro, R. Oostendorp, C. Patillo, M. Soderbom, F. Teal, A. Zeufack,

(2000), "Contract Flexibility and Dispute Resolution in African Manufacturing," *Journal of Development Studies*, 36, 4, 1-37

Blackbourn, D., (1997), *The Long Nineteenth Century*, New York: Oxford University Press.

Blume, L., (2002) "Stigma and Social Control: The Dynamics of Social Norms", mimeo, Department of Economics, Cornell University.

Bowles, S. and H. Gintis, (2002), "Social Capital and Community Governance," *Economic Journal*, 112, 483, 419-436.

Brehm, J. and W. Rahn, (1997), "Individual-Level Evidence for the Causes and Consequences of Social Capital," *American Journal of Political Science*, 41, 3, 999-1023.

Brock, W. and S. Durlauf, (2001a), "Interactions-Based Models," in *Handbook of Econometrics, Vol. 5*, J. Heckman and E. Leamer eds., Amsterdam: North Holland.

Brock, W., and Durlauf, S., (2001b), "Growth Empirics and Reality," *World Bank Economic Review*, 15, 3, 229-272.

Brock, W. and S. Durlauf, (2001c), "Discrete Choice with Social Interactions," *Review of Economic Studies*, 68, 2, 235-260.

Carbonaro, W., (1999), "Opening the Debate: On Closure and Schooling Outcomes," *American Sociological Review*, 64, 682-686.

Carter, M. and M. Castillo, (2003), "An Experimental Approach to Social Capital in South Africa," mimeo, University of Wisconsin.

Carter, M. and M. Castillo, (2004), "Morals, Markets and Mutual Insurance: Using Economic Experiments to Study Recovery from Hurricane Mitch," mimeo, University of Wisconsin.

Carter, M. and J. Maluccio, (2003), "Social Capital and Coping with Economic Shock: An Analysis of Stunting of South African Children," *World Development*, 31, 7, 1147-1163.

Charles, K. and P. Kline, (2002), "Relational Costs and the Production of Social Capital: Evidence from Carpooling," *NBER Working Paper no. 9041.*

Coleman, J., (1988), "Social Capital in the Creation of Human Capital," *American Journal of Sociology*, 94, S95-S121.

Coleman, J., (1990), *The Foundations of Social Theory*. Cambridge: Harvard University Press.

Costa, D. and M. Kahn, (2003a), "Understanding the Decline in American Social Capital, 1953-1998," *Kyklos*, 56, 1, 17-46.

Costa, D. and M. Kahn, (2003b), "Cowards and Heroes: Group Loyalty in the American Civil War," *Quarterly Journal of Economics*, 118, 2, 519-548.

Dasgupta, P., (2000), "Economic Progress and the Idea of Social Capital," in *Social Capital: A Multifaceted Perspective*, P. Dasgupta and I. Seragilden eds., Washington DC: World Bank.

Dasgupta, P., (2002), "Social Capital and Economic Performance: Analytics," mimeo, Faculty of Economics, University of Cambridge.

Dasgupta, P. and I. Serageldin, (2000), *Social Capital: A Multifaceted Perspective*, Washington DC: The World Bank.

Desdoigts, A., (1999), "Patterns of Economic Development and the Formation of Clubs," *Journal of Economic Growth*, 4, 3, 305-330.

DiPasquale, D. and E. Glaeser, (1999), "Incentives and Social Capital: Are Homeowners Better Citizens?," *Journal of Urban Economics*, 45, 2, 354-384.

Durlauf, S., (1999), "The Case "Against" Social Capital," *Focus*, 20, 1-4.

Durlauf, S., (2000), "Econometric Analysis and the Study of Economic Growth: A Skeptical Perspective," in *Macroeconomics and the Real World*, R. Backhouse and A. Salanti, eds., Oxford: Oxford University Press, 2000.

Durlauf, S., (2001), "A Framework for the Study of Individual Behavior and Social Interactions," *Sociological Methodology*, 31, 47-87.

Durlauf, S., (2002a), "The Memberships Theory of Poverty: The Role of Group Affiliations In Determining Socioeconomic Outcomes," in *Understanding Poverty in America*, S. Danziger and R. Haveman, eds., Harvard University Press, 2002.

Durlauf, S., (2002b), "Bowling Alone: A Review Essay," *Journal of Economic Behavior and Organization*, 47, 3, 259-273.

Durlauf, S., (2002c), "On the Empirics of Social Capital," *Economic Journal*, 112, 483, 459-479.

Durlauf, S. and P. Johnson, (1995), "Multiple Regimes and Cross-Country Growth Behavior," *Journal of Applied Econometrics*, 10, 365-384

Durlauf, S., A. Kourtellos, and A. Minkin, (2001), "The Local Solow Growth Model," *European Economic Review*, 45, 928-940, 2001.

Durlauf, S. and D. Quah, (1999), "The New Empirics of Economic Growth," with D. Quah, in *Handbook of Macroeconomics*, J. Taylor and M. Woodford eds., North-Holland.

Durlauf, S. and H. P. Young, (2001), "The New Social Economics," in *Social Dynamics*, S. Durlauf and H. P. Young eds., Cambridge: MIT Press.

Easterly, W. and R. Levine, (1997), "Africa's Growth Tragedy: Politics and Ethnic Divisions," *Quarterly Journal of Economics*, 112, 1203-50.

Ensminger, J. (1992), *Making a Market: The Institutional Transformation of an African Society*, Cambridge University Press, New York.

Fafchamps, M., (1996), "The Enforcement of Commercial Contracts in Ghana," *World Development*, 24, 3, 427-448.

Fafchamps, M., (2002), "Spontaneous Market Emergence", *Topics in Theoretical Economics*, 2, 1, Article 2, Berkeley Electronic Press at www.bepress.com.

Fafchamps, M., (2003), "Ethnicity and Networks in African Trade," *Contributions to Economic Analysis and Policy,* 2, 1, article 14, Berkeley Electronic Press at www.bepress.com.

Fafchamps, M., (2004), *Market Institutions in Sub-Saharan Africa*, Cambridge: MIT Press.

Fafchamps, M. and S. Lund, (2003), "Risk Sharing Networks in Rural Philippines," *Journal of Development Economics*, 71, 261-287.

Fafchamps, M. and B. Minten, (1999), "Relationships and Traders in Madagascar", *Journal of Development Studies*, 35, 6, 1-35.

Fafchamps, M. and B. Minten, (2001), "Social Capital and Agricultural Trade," *Ameircan Journal of Agricultural Economics*, 83, 3, 680-685.

Fafchamps, M. and B. Minten, (2002), "Returns to Social Network Capital Among Traders," *Oxford Economic Papers*, 54, 173-206.

Fernandez, C., E. Ley, and M. Steel, (2001), "Model Uncertainty in Cross-Country Growth Regressions," *Journal of Applied Econometrics*, 16, 5, 563-576.

Fernandez, R., E. Castilla, and P. Moore, (2000), "Social Capital at Work: Employment at a Phone Center," *American Journal of Sociology*, 105, 1288-1356

Fershtman, C. and U. Gneezy, (2001), "Discrimination in a Segmented Society: An Experimental Approach," *Quarterly Journal of Economics*, 116, 1, 351-77.

Frank, K., and J. Yasumoto, (1996), "Linking Action to Social Structure within a System: Social Capital within and between Subgroups," *American Journal of Sociology*, 104, 3, 642-686.

Fudenberg, D. and E.. Maskin (1996), "The Folk Theorem in Repeated Games with Discounting," *Econometrica*, 54, 533-556

Fukuyama, F. (1997), "Social Capital," Tanner Lecture on Human Values.

Furstenberg, F. and M. Hughes, (1995), "Social Capital and Successful Development Among At-Risk Youth," *Journal of Marriage and the Family*, 57, 580-592.

Geertz, C., H. Geertz, and L. Rosen, (1979), *Meaning and Order in Moroccan Society*, Cambridge: Cambridge University Press.

Glaeser, E., D. Laibson, J. Scheinkman and C. Soutter, (2000), "Measuring Trust," *Quarterly Journal of Economics*, CXV, 811-846.

Glaeser, E., D. Laibson, and B. Sacerdote, (2002), "An Economic Approach to Social Capital," *Economic Journal*, 112, 483, 437-458.

Goldin, C. and L. Katz, (1999), "Human Capital and Social Capital: the Rise of Secondary Schooling in America. 1910 to 1940," *Journal of Interdisciplinary History*, 29, 683-723.

Granovetter, M., (1975), *Getting a Job: A Study of Contacts and Careers*, Chicago: University of Chicago Press, Chicago; 2nd edition 1995

Granovetter, M., (1985), "Economic Action and Social Structure: The Problem of Embeddedness," *American Journal of Sociology*, 91, 3, 481-510.

Granovetter, M. (1995) "The Economic Sociology of Firms and Entrepreneurs", in *The Economic Sociology of Immigration: Essays on Networks, Ethnicity, and Entrepreneurship*, Alejandro Portes, ed., New York: Russell Sage Foundation, 128-165.

Greif, A., (1993), "Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition," *American Economic Review*, 83, 3, 525-548

Grootaert, C., (2000), "Social Capital, Household Welfare, and Poverty in Indonesia," mimeo, World Bank.

C. Grootaert and T. van Bastelear, (2002), *The Role of Social Capital in Development: An Empirical Assessment*, Cambridge: Cambridge University Press.

Gugerty, M. and M. Kremer, (2002), "The Impact of Development Assistance on Social Capital: Evidence from Kenya," in *The Role of Social Capital in Development: An*

Empirical Assessment, C. Grootaert, and T. van Bastelear, eds., Cambridge: Cambridge University Press.

Guiso, L., P. Sapienza, and L. Zingales, (2002), "The Role of Social Capital in Financial Development," *George J. Stigler Center for the Study of the Economcy and the State Working Paper no. 173.*

Hagan, J., R. MacMillan, and B. Wheaton, (1996), "New Kid in Town: Social Capital and the Life Course Effects of Family Migration on Children," *American Sociological Review*, 61, 368-385.

Hagan, J. and B. McCarthy, (1995), "Getting Into Street Crime: The Structure and Process of Criminal Embeddedness," *Social Science Research*, 24, 63-95.

Hallinan, M., and W. Kubitschek, (1999), "Conceptualizing and Measuring School Social Networks: Comment on Morgan and Sorenson," *American Sociological Review*, 64, 687-693.

Hayek, F. A., (1945), "The Use of Knowledge in Society", *American Economic Review*, 35, 4, 519-530

Helliwell, J., (1996), "Economic Growth and Social Capital in Asia," in *The Asia Pacific Region in the Global Economy: A Canadian Perspective*, R. Harris, Richard, ed., Calgary: University of Calgary Press, 1996.

Helliwell, J. and R. Putnam, (2000), "Economic Growth and Social Capital in Italy," in *Social Capital: A Multifaceted Perspective*, P. Dasgupta and I. Seragilden eds., Washington DC: World Bank, 253-266.

Hofferth, S., J. Boisjoly, and G. Duncan, (1999), "The Development of Social Capital," *Rationality and Society*, 11, 1, 79-110.

Howitt, P. and D. Mayer-Foulkes, (2002), "Technological Innovation, Implementation, and Stagnation: A Schumpeterian Theory of Convergence Clubs," *National Bureau of Economic Research Working Paper no. 9104.*

Isham, J., (2002), "The Effect of Social Capital on Fertilizer Adoption: Evidence from Rural Tanzania," *Journal of African Economics*, 11, 1, 39-60.

Jimenez, E. and Y. Sawada, (1999), "Do Community-Managed Schools Work? An Evaluation of El Salvador's EDUCO Program," *World Bank Economic Review*, 13, 3, 415-41.

Johnson, S., J. McMillan, and C. Woodruff, (2000), "Entrepreneurs and the Ordering of Institutional Reform: Poland, Slovakia, Romania, Russia and Ukraine Compared," *Economics of Transition*, 8, 1, 1-36

Kandori, M., (1992), "Social Norms and Community Enforcement", *Review of Economic Studies*, 59, 63-80.

Knack, S. and Keefer, P., (1997), "Does Social Capital Have an Economic Impact? A Cross-Country Investigation," *Quarterly Journal of Economics*, 112, 1252-1288.

Kranton, R., (1996), "Reciprocal Exchange: A Self-Sustaining System", *American Economic Review*, 86, 4, 830-851

Krishna, A., (2001), "Moving from the Stock of Social Capital to the Flow of Benefits: The Role of Agency," *World Development*, 29, 925-943.

Krishna, A. and N. Uphoff, (1999), "Mapping and Measuring Social Capital: A Conceptual and Empirical Study of Collective Action for Conserving and Developing Watersheds in Rajastan, India, *World Bank, Social Capital Initiative Working Paper no. 13.*

LaPorta, R., F. Lopez-de-Silanes, A. Shleifer, and R. Vishny, (1997), "Trust in Large Organizations," *American Economic Review*, 87, 333-338.

Lee, S. and M. Brinton, (1996), "Elite Education and Social Capital: The Case of South Korea," *Sociology of Education*, 69, 177-192.

Lin, N., (2001), *Social Capital*, Cambridge: Cambridge University Press.

Lochner, K., I. Kawachi, R. Brennan, and S. Buka, (2003), "Social Capital and Neighborhood Mortality Rates in Chicago," *Social Science and Medicine*, 56, 1797-1805.

Loury, G., (1977), "A Dynamic Theory of Racial Income Differences," in *Women, Minorities, and Employment Discrimination*, P. Wallace and A. LeMund, eds., Lexington: Lexington Books.

Loury, G., (1981), "Intergenerational Transfers and the Distribution of Earnings," *Econometrica*, 49, 843-867.

Maluccio, J., L. Haddad, and J. May, (2001), "Social Capital and Household Welfare in South Africa," *Journal of Development Studies*, 36, 6, 54-81.

Manski, C., (1993), "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 60, 531-542.

Manski, C., (2000), "Economic Analysis of Social Interactions," *Journal of Economic Perspectives*, 14, 114-136.

Manski, C., (2002), "Identification of Decision Rules in Experiments on Simple Games of Proposal and Response," *European Economic Review*, 46, 880-891.

McMillan, J. and C. Woodruff, (August 2000), "Private Order Under Dysfunctional Public Order," *Michigan Law Review,* 98, 8, 2421-2458.

McNeal, R., (1999), "Parental Involvement as Social Capital: Differential Effectiveness on Science Achievement, Truancy, and Dropping Out," *Social Forces*, 78, 1, 117-144.

Miguel, E., P. Gertler, and D. Levine, (2001), "Did Industrialization Destroy Social Capital in Indonesia?," mimeo, Department of Economics, University of California at Berkeley.

Montgomery, J., (1991), "Social Networks and Labor-Market Outcomes: Toward an Economic Analysis", *American Economic Review*, 81, 5, 1408-1418.

Morgan, S. and A. Sorenson, (1999a), "Parental Networks, Social Closure, and Mathematics Learning: A Test of Coleman's Social Capital Explanation of School Effects," *American Sociological Review*, 64, 661-681.

Morgan, S. and A. Sorenson, (1999b), "Theory, Measurement, and Specification Issues in Models of Network Effects on Learning: Reply to Carbonaro and to Hallinan and Kubitschek," *American Sociological Review*, 64, 693-700.

Narayan, D. and L. Pritchett, (1999), "Cents and Sociability: Household Income and Social Capital in Rural Tanzania," Economic Development and Cultural Change, 47, 4, 871-897.

North, D., (1973), *The Rise of the Western World*, Cambridge: Cambridge University Press.

North, D., (1990), *Institutions, Institutional Change, and Economic Performance*, Cambridge: Cambridge University Press.

North, D., (2001), "Comments", in *Communities and Markets in Economic Development*, 403-8, M. Aoki and Y. Hayami eds., Oxford: Oxford University Press.

Oliver, J., (1999), "The Effects of Metropolitan Economic Segregation on Local Civic Participation," *American Journal of Political Science*, 43, 186-212.

Ostrom, E., (2000), "Social Capital: A Fad or Fundamental Concept?," in *Social Capital: A Multifaceted Perspective*, P. Dasgupta and I. Seragilden eds., Washington DC: World Bank.

Ostrom, E. and T.Ahn, (2002), "A Social Science Perspective on Social Capital: Social Capital and Collective Action," mimeo, Workshop in Political Theory ad Policy Analysis, Indiana University.

Paldam, M. and G. Svendsen, (2000), "An Essay on Social Capital: Looking for Smoke Behind the Fire," *European Journal of Political Economy*, 16, 339-366.

Palloni, A., D. Massey, M. Ceballos, K. Espinosa, and M. Spittel, (2001), "Social Capital and International Migration: A Test Using Information on Family Networks," *American Journal of Sociology*, 106, 1262-1298.

Parcel, T. and E. Menaghan, (1993), "Family Social Capital and Children's Behavioral Outcomes," *Social Psychology Quarterly*, 56, 2, 120-135.

Pargal, S., M. Huq, and D. Gilligan, (1999), "Social Capital in Solid Waste Management: Evidence from Dhaka, Bangladesh," *World Bank, Social Capital Initiative Working Paper no. 16.*

Paxton, P., (1999), "Is Social Capital Declining? A Multiple Indicator Assessment," *American Journal of Sociology*, 105, 88-127.

Paxton, P. (2002), "Social Capital and Democracy: An Interdependent Relationship," *American Sociological Review*, 67, 254-277.

Platteau, J.-P., (1994a), "Behind the Market Stage Where Real Societies Exist: Part I - The Role of Public and Private Order Institutions," *Journal of Development Studies*, 30, 3, 533-577.

Platteau, J.-P., (1994b), "Behind the Market Stage Where Real Societies Exist: Part II - The Role of Moral Norms," *Journal of Development Studies*, 30, 3, 533-577.

Platteau, J. P. and E. Seki (2002), "Community Arrangements to Overcome Market Failure: Pooling Groups in Japanese Fisheries", in *Communities and Markets in Economic Development*, M. Aoki and Y. Hayami eds., Oxford: Oxford University Press.

Poewe, K. (1989) *Religion, Kinship, and Economy in Luapula, Zambia*, The Edwin Mellen Press, Lewinston.

Portes, A., (1998), "Social Capital: Its Origins and Application in Modern Sociology." *Annual Review of Sociology*, 1-14.

Portes, A., (2000), "The Two Meanings of Social Capital." *Sociological Forum,*15, 1,:1-12.

Putnam, R. with R. Leonardi and R. Nanetti, (1993), *Making Democracy Work: Civic Traditions in Modern Italy*, Princeton: Princeton University Press.

Putnam, R., (1995), "Tuning In, Tuning Out: The Strange Disappearance of Social Capital in America," *PS: Political Science & Politics*, December, 664-683.

Putnam, R. (2000), *Bowling Alone*, New York: Simon and Schuster.

Rahn, W. and T Rudolph, (2002), "A Multilevel Model of Trust in Local Government," mimeo, University of Minnesota.

Rauch, J. and A. Casella, (2001) "Overcoming Informational Barriers to International Resource Allocation: Prices and Group Ties," *Economic Journal*, 113, 484, 21-42.

Robison, L. and M. Siles, (1999),"Social Capital and House Income Distributions in the United States: 1980, 1990," *Journal of Socio-Economics*, 28, 43-93.

Roemer. J., (1998), *Equality of Opportunity*, Cambridge: Harvard University Press.

Routledge, B. and J. von Amsburg, (2003), "Social Capital and Growth," *Journal of Monetary Economics*. 50, 1, 167-194.

Sampson, R., Morenoff, J., and Earls, F., (1999), "Beyond Social Capital: Collective Efficacy for Children," *American Sociological Review*, 64, 633-660.

Sandefur, G., A. Meier, and P. Hernandez, (1999), "Families, Social Capital, and Educational Continuation," mimeo, Department of Sociology, University of Wisconsin at Madison.

Sandefur, R. and E. Laumann, (1998), "A Paradigm for Social Capital," *Rationality and Society*, 10, 4, 481-501.

Skocpol, T., (1996), "Unravelling from Above," *The American Prospect*, March-April, 20-25.

Skocpol, T., (2003), *Diminished Democracy: From Membership to Management in American Civic Life*, Norman: University of Oklahoma Press.

Somanathan, E. and R. Rubin, (2004), "The Evolution of Honesty," *Journal of Economic Behavior and Organization*, 54, 1-17.

Sun, Y., (1999), "The Contextual Effects of Community Social Capital on Academic Performance," *Social Science Research*, 28, 4, 403-426.

Tarrow, S., (1996), "Making Social Science Work Across Space and Time: A Critical Reflection on Robert Putnam's *Making Democracy Work*," *American Political Science Review*, 90, 2, 389-397.

Taylor, C., (2000), "The Old-Boy Network and the Young-Gun Effect," *International Economic Review*, 41, 4, 871-91

Teachman, J., K. Paasch, and K. Carver, (1997), "Social Capital and the Generation of Human Capital," *Social Forces*, 75, 4, 1-17.

Temple, J., (1999), "The New Growth Evidence," *Journal of Economic Literature*, 37, 112-156.

Temple, J., (2000), "Growth Regressions and What the Textbooks Don't Tell You," *Bulletin of Economic Research*, 14, 395-426.

Varughese, G. and Ostrom, E., (2001), "The Contested Role of Heterogeneity in Collective Action: Some Evidence from Community Forestry in Nepal," *World Development*, 29, 747-765.

Wade, R., (1987), "The Management of Common Property Resources: Finding a Cooperative Solution", *World Bank Research Observer*, 2(2): 219-34.

Wade, R., (1988), "The Management of Irrigation Systems: How to Evoke Trust and Avoid Prisoners' Dilemma", *World Development*, 16(4): 489-500.

Woolcock, M., (1998), "Social Capital and Economic Development: Toward a Synthesis and Policy Framework," *Theory and Society*, 27, 151-208.

Zak, P. and S. Knack, (2001), "Trust and Growth," *Economic Journal*, 111, 295-321.

THE EFFECT OF ECONOMIC GROWTH ON SOCIAL STRUCTURES [1]

François Bourguignon
The World Bank

October 23, 2004

[1] Prepared for the Handbook of Economic Growth edited by Philippe Aghion and Steve Durlauf. This paper was finished, with great difficulty, after I joined the World Bank as Chief Economist in the summer 2003. I am grateful to Philippe Aghion for useful comments. I thank for their patient and efficient help Jean-Jacques Dethier and Victoria Levin. Views expressed here are essentially personal and do not engage the World Bank in any manner.

INTRODUCTION

If economic growth actually resembled the 'extended reproduction' coined by Marx and implicit in the steady state regimes of many contemporary growth models, one would not expect growth to have major social consequences. All economic magnitudes, including the standards of living of individuals or social groups, would be kept in the same proportion to each other so that only the scale of the economy would be changing over time.

Of course, economic growth is something more than a mere uniform change of scale of economic magnitudes. For a host of reasons, it is in the very nature of growth to modify economic structures and, because of this, to affect social structures and social relations. For instance, growth may modify the sectoral structure of an economy leading firms in one sector to close down and firms in other sectors to be created or expand. Growth modifies the structure of prices, thus affecting the standard of living of households in a way that depends on their consumption preferences. In other cases, growth will call on some particular skills, increasing the remuneration of those endowed with those skills and also, possibly, their decision-making power within society. Finally, growth may reduce the availability of public goods like clean air or water, requiring public intervention in order to maintain the adequate supply of environmental goods. In all these cases, it is not only the economic structure – i.e., the relative importance of sectors, labor skills, remuneration of factors, and size of the public sector – that may be modified by growth. It is also the whole social structure, that is the relative weight of socio-economic groups or the way in which individuals define themselves with respect to the rest of the society, that is affected. As a consequence, social relations that govern how individuals in a society interact with each other through explicit or implicit rules may also be modified by economic growth and may in turn affect the growth process itself.

Rough evidence of such changes is provided by simple comparisons of economic and social structures and institutions across countries which have reached different levels of development. At the risk of caricaturing, it is sufficient to compare poor sub-Saharan

African countries today with some highly developed countries in Europe, North America or in the South Pacific. At one end of the spectrum, one observes largely rural societies dominated by household farms, few wage workers except in the limited urban sector, social protection ensured by an extended family system and a relatively small public sector often controlled by an unstable oligarchy. At the other end, one finds almost exclusively democratic urban societies with salaried employment and private ownership of capital as the main economic organization, with sophisticated redistribution system run by governments the size of which is 3 to 4 times that observed in poorer countries.

It is tempting to attribute all of these differences to economic growth and to expect that growth in the poorest countries will progressively make them comparable to developed countries today. Unfortunately, things are not so simple. In particular, it is clear that differences in economic and social structures and institutions cause differences in the pace and structure of economic growth as much as they are caused by it. But, it is also the case that some other factors may be influencing simultaneously both the process of economic growth and social structures and institutions. To take an example, a longer life expectancy due to technical advances in the field of health is likely to modify social structures through the aging of the population but it is also likely to modify economic behavior and the growth process, for instance because higher saving rates are rendered necessary by the prospect of longer periods of inactivity. In turn, this effect on economic growth rates and on the level of development may affect social structures and institutions by changing the weight of particular sectors in the economy.

The effects of economic growth on social structures are more complex than what the reduced form regressions found in recent literature may suggest. The relationships are likely to be non-linear (as hypothesized, for instance, by Kuznets for income inequality) and to depend on several country characteristics, including policy and institutional variables. This chapter argues that simple statistical methods are unable to identify these forces and these interactions, and that the limited number of observations available is a serious hindrance for this identification. Under these conditions, the only methodological approach able to identify the social consequences of economic growth is of a structural

nature. It first requires establishing hypotheses about the channels through which economic growth may affect the social structures under analysis. These hypotheses should then be empirically tested, provided of course that the data necessary to do so are available.

The purpose of this chapter is to examine, among the changes in social structures that may be observed along the growth path of a country or when comparing countries at different levels of development, what may be considered to be the *direct* consequences of economic growth. Other changes, thus, have to be considered as autonomous or possibly caused by factors or initial conditions that may have also affected growth but that have only an *indirect* relation to growth. An important example of such autonomous changes would be technological progress.

As noted above, some of these direct social consequences of growth may affect the pace and the structure of future growth and thus feed back into themselves through various channels. For instance, growth may under some circumstances generate more inequality in the distribution of economic resources, this increased inequality in turn affecting the dynamics of the economy. This chapter focuses on that part of the circular argument that goes from growth to social structures and ignores the other side of the circle. Such a choice is made for reasons of analytical expediency. It turns out that the economic mechanisms that lie behind the two parts of the circle are quite different; it would be too ambitious a task to deal with them simultaneously. Readers interested in the effect of social structures and institutions on growth should refer to other chapters in this Handbook.

The social consequences of growth may be of diverse nature. A natural distinction to be made is between the consequences of growth for ‘social structures’ and the consequences for ‘social relations' and 'social institutions'. As suggested by Kuznets (1966, p. 157-8), c*hanges in social structures* have to be understood essentially as the differential effects of economic growth on predetermined social groups. For example, urban skilled workers may benefit more from economic growth at some stage than unskilled workers, rentiers

more than farmers, men than women, or young people than older people. Growth may also affect the size of those various groups. Some substantial proportion of people migrate from the countryside to the cities under the pressure of urban growth, or more people may be willing to acquire a secondary or tertiary educational level. In both cases, social distances between individuals are modified. Changes in *social institutions* may result from changes in these social structures or from autonomous forces. For instance, changing the weight of specific socio-economic groups within society modifies the dominant mode of social relations, and changing the economic distance between individuals may modify the way they interact.

This chapter concentrates on the consequences of growth for social structures rather than on social institutions or relations. Its ambition is to identify the role played by economic growth in observed changes in social structures and to disentangle it from other factors. Casual comparison of social structures in developed and developing countries shows obvious and enormous differences. However, just because these two country groups differ by their mean income level does not imply that observed social differences must be exclusively attributed to economic growth per se. There may be many other reasons for these differences. In particular, it is possible that initial or historical conditions are responsible for some specific social evolution and for a particular growth path in a given country. It is also possible that exogenous forces, such as technical progress, have a direct specific impact on social structures, on the one hand, and on the pace and structure of economic growth, on the other. Analyzing the social consequences of growth consists of trying to isolate somehow the pure 'income effect' in the evolution of social structures.

This chapter is organized in three parts. The first part introduces the topic by examining the nature of the statistical relationships existing between social indicators and development across countries and/or across periods. It illustrates the differences in social structures associated with differences in income, but it also shows the difficulty of obtaining precise estimates of the size of the 'income effect' from this kind of evidence and the need to rely on more structural analyses. The second part reviews theoretical models of the effect of economic growth on social structures, with an emphasis on

several dimensions of social differentiation and on economic inequality. Finally, the third part focuses on the empirical evidence in support of this structural view of the consequences of growth for social structures.

1. Statistical Relationships between Growth and Social Structures

One may think of literally thousands of aggregate characteristics of societies showing extremely high degrees of correlation with indicators of economic development, either when comparing different countries at different levels of development or when analyzing the evolution of a single country over time. Collecting all existing results of this nature in the economic and non-economic literature is beyond the scope of this chapter.[2] Moreover, it is not clear how informative these correlations are from the point of view of causality. This section aims at showing that even the most sophisticated statistical techniques for the analysis of the relationship between socio-economic indicators and the level of development are unlikely to permit identifying the desired causality link between them. Given the available evidence, identifying that link requires dealing implicitly or explicitly with structural models, rather than with the reduced form models behind correlation analysis, whatever the degree of technical sophistication of that analysis.

As an example of the correlation approach to the consequences of growth, Table 1 shows the relationship between the level of economic development and a few indicators that very roughly describe changes in societies' economic and social structure generally associated with economic growth. As it will be seen later in this chapter, these indicators describe important channels through which growth and development may modify social structures. They include the size of the government, the level of urbanization, education, health, demographic patterns, labor force participation, gender differences and income inequality. The first three columns of the table report the results of a simple regression of

[2] For an early comprehensive attempt of this type, see Adelman and Morris (1967) who argue that development is a complex multi-causal process explained by many interactions between social, economic, political and institutional variables and use factor analysis to reduce the large number of explanatory variables into a small number of key categories. Zhang, Johnson, Resnick and Robinson (2004) present a typology of development strategies applying the same technique to Sub-Saharan Africa.

these indicators on GDP per capita expressed in 2000 US dollars after correction for purchasing power parity. The first two columns are based on pure cross-country observations – observed country means for the 1970s and the 1990s - whereas the third one is based on a pooling of all data available across countries and years during the period 1960-2002.[3]

It can be checked there that, with two exceptions, all indicators appear to be significantly and strongly correlated with economic growth. For instance, focusing on the pooled regression, the GDP share of public expenditures is shown to increase by 0.5 percentage point when GDP increases by $ 1,000 (thus confirming 'Wagner's law') although this coefficient is not statistically significant for the 1970 cross-section. Likewise, the urbanization rate is shown to increase by 0.3 per cent and the literacy rates by 3 to 4 percentage points in presence of the same increase in income per capita, whereas fertility decreases by 0.15 children; the 1970 cross-section shows slightly different results in all of these cases. As a final example, income inequality and female gender bias appear to be significantly and negatively correlated with growth. In the case of the former, a parabolic regression exhibits the familiar inverted U-shape introduced by Kuznets some 50 years ago [4]– with a non-significant result occurring with the 1990 cross-section.

<Table 1 around here>

All these results are interesting. Yet, there are various reasons to think that simple regression on a cross-section of countries or even on a pool of cross-section time-series observations is a very crude approach to identifying the consequences of growth. On the one hand, the existence of a correlation does not say much about the causality link between two variables. Causality may be direct in either one direction, or possibly in both. It may also be indirect and simply reflect the fact that the two variables under scrutiny are both related to a common set of other variables. On the other hand, GDP per

[3] All data are from World Bank's SIMA database (World Bank 2003).

[4] On the basis of historical data, Kuznets (1955) proposed the hypothesis that income inequality tended to increase in a first stage of economic development and to fall in a second one. See the discussion in section 2 below.

capita tends to increase more or less regularly over time so that there may be a confusion between its effect on socio-economic indicators and that of other variables with a comparable time trend.

Alternative econometric specifications permit taking into account some of the preceding points. At the same time, however, they often modify the order of magnitude and the significance of the preceding relationships. In a few instances, they even modify their direction.

The next four columns of table1 show estimates of the growth sensitivity of socio-economic indicators obtained with alternative econometric specifications. In all cases, the sample is obtained by pooling country data over various years in the periods 1960-2002. In column 4, a set of year dummy variables is added to the regression. This accounts for the fact that socio-economic indicators might evolve over time under the influence of some common factors independent of national economic growth. In column 5, it is a set of country dummy variables that is introduced so as to control for 'fixed effects' or, in other words, the effect of largely unobserved fixed country characteristics that might affect both the original level of GDP per capita and that of the indicator under scrutiny. The corresponding estimate of growth sensitivities thus abstracts from differences in country means and takes into account only differences in the average time behavior of GDP per capita and socio-economic indicators across countries. Column 6 combines both approaches by abstracting from differences in country means as well as from an exogenous non-linear time trend common to all countries. Finally the estimates in column 7 are obtained by running the simple regressions of socio-economic indicators on GDP per capita in decadal differences.

Adding a common non-linear time trend to the original simple model does not modify the growth sensitivity of the socio-economic indicators in a significant way. More substantial changes are obtained when fixed country effects are introduced. As could be expected, growth sensitivity generally falls when cross-country differences are ignored, or more exactly when cross-country differences are attributed to fixed characteristics that include,

inter alia, initial development levels. The effect of growth on the urbanization rate, the literacy rate or life expectancy is divided by about 2 The growth sensitivity of the GDP share of public expenditures becomes non-significant, the same being true of income inequality, both with the linear and with the parabolic model. The only exception is labor force participation of women, the effect of growth on which tends to increase when controlling for fixed effects. Changes with respect to simple estimates are still bigger when fixed effects are introduced both for countries and years. In some cases – as for instance with fertility and gender life expectancy differential- the sign of growth sensitivity is even reversed.

Of course, such a correction of the original estimates may well be excessive. Adding a time trend is certainly bound to reduce growth sensitivity estimates, especially when estimation abstracts from cross-country differences. The time trend is likely to pick up those changes in the indicators which are independent from country specific economic growth. Yet, results obtained with that method are not always very convincing. In particular, that fertility would significantly increase as a response to growth once independent forces are taken into account, as shown in the last two columns of table 1, seems to be in contradiction with the intuition of most demographers.[5] The results shown in table 1 are likely to mask some heterogeneity among countries with respect to the drop in fertility that is not dependent on economic growth.

The estimates reported in the last column of table 1 confirm the preceding results. Restricting the analysis to correlation in decadal differences overall shows the same order of magnitude for the growth sensitivity of socio-economic indicators, but also makes those sensitivities often non-significant.[6] In comparison with simple regressions and correlations, estimates based on differences or on fixed effects thus suggest that the evolution of the few general socio-economic indicators considered in table 1 probably obeys other forces in addition to economic growth, or that the effect of growth is less

[5] See for instance Easterlin (1996) chapter 8, and Lee (2003).
[6] The difference in T-statistics between columns (6) and (7) is mostly due to the fact that the regression in column (7) relies on fewer observations. Note, however, that ignoring possible correlation in the residuals of the regressions behind column (6) for contiguous years may tend to a gross underestimation of the variance of the estimates.

simple than implicitly assumed in these statistical models. In particular, it is quite possible that the effects of growth on socio-economic indicators are strongly heterogeneous across countries. Identifying that heterogeneity or the forces other than economic growth that affect the evolution of socio-economic indicators is thus necessary in order to identify the true social consequences of economic growth. But the econometric approach illustrated in this section is unlikely to meet that objective.

Taking into account this heterogeneity across countries and going beyond the simple statistical techniques used to produce the results of table 1 meets a fundamental constraint: the limited number of observations. Estimating the preceding model country by country on a time series basis would certainly permit to fully account for country specificity. But it would only inform on the consequences of growth at a particular stage of the development process of a given country. On the other hand, taking into account observed heterogeneity by interacting growth rates or development levels with a host of country characteristics and policy variables and by introducing non-linearity is also bound to run into too few degrees of freedom. Social consequences of growth take time to show up, so that the informational content of annual time series is not proportional to the length of these series. In table 1, one can see that there is little difference between the last two columns even though column (6) relies on full annual series whereas column (7) uses only decadal differences. This means that the information that matters is not year-to-year fluctuation but 'episodes' of growth characterized by uniformly high, moderate or low growth rates. If this is the case, then available data may make it difficult to estimate with satisfactory precision the observed heterogeneity in the consequences of growth.[7]

In summary, the analysis in this section suggests that other forces than economic growth are behind the time evolution of most socio-economic indicators, even though the absolute value of simple correlation coefficients is often very high. It is also possible that

[7] Another approach to estimating the growth sensitivity of social indicators would be to estimate jointly the difference equation in column (7) and the level equations in columns (1) and (2). Resulting estimates would simply be midway. Working with annual series, it would also be possible to explicitly introduce some lag in the effect of growth on social indicators and to use GMM-based Arellano-Bond or Blundell-Bond 'system' estimates (Arellano and Bond 1991, Blundell and Bond 1998) instead of the fixed effect model (6). Given the proximity of the estimates in columns (6) and (7), the lag is likely to be quite long and the overall sensitivity not very different from the estimates shown in these columns.

the effect of economic growth on these indicators and the social features they describe is too complex to be described by simple regression analysis. In particular, the relationship may be non-linear – as hypothesized for instance by Kuznets for income inequality – or it may depend on specific country characteristics, including policy and institutional variables. There is no simple statistical method to identify *a priori* these other forces or these interactions, and limited observations may also be a serious hindrance for this identification.

Under these conditions, it is likely that the only methodological approach able to help identify and understand the consequences of economic growth is of a 'structural' nature. In other words, what is required is to establish hypotheses on the phenomena that guide the overall evolution of the socio-economic indicators under analysis and to test the corresponding model. This is in stark contrast with the reduced form approach so often found in the literature and illustrated by table 1. Within a structural approach, a theoretical model of the behavior of the socio-economic indicator being studied must first be established on the basis of economic theory. This is what the next section attempts to do for some possible social consequences of growth.

2. THE EFFECT OF ECONOMIC GROWTH ON SOCIAL STRUCTURES : THEORETICAL CONSIDERATIONS

Does economic growth tend to affect people's income and welfare in the same way? Alternatively, does economic growth tend to favor the expansion of some particular socio-economic groups with respect to others? Of course, these two types of changes in social structures are related to each other. It is because economic growth favors urban workers over rural workers that the latter tend to migrate to the cities and the size of the agricultural population declines. Likewise, it may be because, in some circumstances, growth favors skilled work that parents have an incentive to send their kids to school for longer periods and the literacy rate in the labor force tends to rise. In line with the plea for a 'structural approach' to these questions in the preceding section, this section reviews various theoretical models meant to describe the 'distributional' consequences of growth.

It is well-known, since the pioneering work by Chenery and Syrquin (1975) on the structural aspects of growth, that growth favors some specific sectors and therefore specific social groups, mostly those who work in them or consume their products. Thus, when growth occurs, changes take place in the weight of these sectors in the economy and in the weight of these groups in the population. However, it is not necessarily the case that those changes also produce permanent modifications in relative incomes. Indirect mechanisms might partially or completely compensate for these direct effects by spreading them to the rest of the society. Competition on the labor market may spread sectoral effects to the whole economy, for instance, or migration may be a natural response to urban-oriented growth leaving urban-rural income differentials unchanged. Evaluating this chain of effects may need fairly elaborate models, however. Evaluating the effect of growth on social structures thus is more or less straightforward depending on what aspect of social structures is being studied. Given the considerable interest it arose in the literature over the past 30 years or so, this review concentrates on the issue of income inequality, while considering at the same time related dimensions of social differentiation.

The relationship between economic growth and the inequality of the distribution of income and economic resources in general has attracted the interest of economists ever since the classical age of the discipline. More recently, very much interest arose with Kuznets' (1955) observation that, historically, inequality tended to increase in a first stage and then to decrease at a later stage of development. Cross-country analysis undertaken in the early 1970s by Paukert (1973) and Ahluwalia (1976) seemed to confirm that there was indeed an inverted U-shape relationship between inequality and the level of development, as measured by the GDP per capita.

As can be seen in at the bottom of table 1, the cross-country data available circa 1970 seemed indeed in full agreement with Kuznets' hypothesis. Data that became available later did not confirm that feature of the 1970 sample, however, whereas estimations based

on panel data suggested that, in many countries, the evolution of inequality did not fit Kuznets' patterns. In table 1, the coefficient of GDP per capita (y) and its square (y^2) show an inverted U-curve in 1970s, a significant U-shaped curve when these data are pooled together with more recent data, and a non-significant relationship when controlling for fixed effects.

Interestingly enough, the debate about the Kuznets hypothesis gave rise to a renewal of the theoretical literature on the general effects of growth on inequality. That literature also provides a representation of the channels through which economic growth may affect social structures in general. This literature is briefly reviewed in what follows.

The literature on the effect of growth on inequality emphasizes two fundamental channels: sectoral shifts, on the one hand, and factor markets on the other. Both channels have been represented with different types of modeling and have been subject to continuous scrutiny and analytical elaboration. They remain the cornerstones of any analytical approach to the social consequences of growth.

(a) The Sectoral Shift View

The explanation that Kuznets himself gave to the inverted U-curve hypothesis was based on the sectoral shifts away from traditional agriculture that characterizes long-run economic growth. In effect, the model he had in mind was very much along the lines of the classical surplus labor model as formulated in the modern literature by Lewis (1954) and later by Fei and Ranis (1965). There are two sectors in the economy with fixed relative prices and fixed relative incomes. The development process consists of shifting some proportion of the population from one sector to the other. An obvious formalization of this model is as follows. Let y_i be the fixed income level in sector i, and n_i the share of the population in that sector. Let sector *2* be traditional agriculture and suppose that income in that sector is smaller than in the 'modern' sector, labeled *1* – i.e. $y_1 > y_2$. Long-run growth in that model is then essentially described by the increase in the proportion of the population employed in the modern sector, n_1, for fixed income levels y_1 and y_2. Such

a process may for instance be explained by some capital accumulation taking place in sector *1* and some labor-market imperfection preventing labor remuneration to equalize in the two sectors.

This simple representation of the process of economic growth has obvious implications for social structures in general. Everything depends on the interpretation given to the two sectors *1* and *2*. If sector *2* is indeed assimilated with traditional agriculture, then the drop in n_2, and the consequent increase in n_1, implies altogether an increase in urbanization and all social transformations that may possibly accompany it, like lower fertility, higher school enrolment, higher crime rate, etc… But the dichotomy between sectors *1* and *2* may also represent manufacturing versus services, formal versus informal or high versus low technology. In each case, growth comes with a more or less rapid modification in the structure of society in a particular dimension.

This framework may be easily extended so as to represent the evolution of income inequality within society. Following Robinson (1976), let income inequality be measured by the variance of the logarithm of income.[8] Thus, denote V_i the variance in sector *i* and assume that this variance is constant. Total income inequality in the economy is then given by:

$$V = n_1.V_1 + (1-n_1).V_2 + n_1.Log^2 y_1 + (1-n_1).Log^2 y_2 - \left[n_1.Log\ y_1 + (1-n_1).Log\ y_2\right]^2 \quad (1)$$

where use is made of the fact that $n_1 + n_2 = 1$. As before, the development level of the economy is fully described by the proportion of people in sector *1*, n_1. If it is assumed that $V_1 > V_2$, then total inequality in (1) is a parabolic function of n_1. Under some plausible conditions on the values of V_1 , V_2, y_1 and y_2, total inequality may thus go up and then down as observed by Kuznets on some historical data.

8 Knight (1976) and Fields (1979) use the same framework but different income inequality measures. Anand and Kanbur (1993) present a more general version of this model where the analysis is conducted in terms of the full distribution of income in both sectors and in the whole economy rather than on a specific summary inequality measure.

Yet, this result on inequality, and more generally the fact that this model represents the social consequences of growth through a single parameter, n_1, must be taken with very much care. First, as n_1 is bounded, it may well be the case that inequality will be increasing or decreasing monotonically throughout the development process. Depending on the various parameters of the model, the time profile of inequality may be extremely flat or, on the contrary, have a sizable slope. Second, representing growth through a mere sectoral shift of the population may seem overly simplistic and restrictive. Assuming that income in the two sectors of the economy does not change along with growth is equivalent to assuming that markets are imperfectly competitive or that compensating phenomena are at work. Practically, more people in sector *1* could lower the relative price of that sector's output. This might reduce the initial level of inequality between the two groups of workers and may be enough to prevent inequality among individuals to go up in the first stage of the process just described. Likewise, it is restrictive to assume that migration from one sector to the other is distribution neutral. A change in n_1 is likely to modify both V_1 and V_2. The direction of that change will depend on whether migration concerns the least well off people in sector *2* or people in the middle of the income scale.

In short, representing economic growth trough a simple sectoral shift parameter appears unsatisfactory to account for the effect of growth on social structures, except maybe in very particular cases where inequality across socio-economic groups may indeed be considered as constant. In general, however, the implicit fixed price assumption in the sectoral shift model seems unduly restrictive.

Several authors have proposed extensions of the preceding model for the analysis of the evolution of inequality among individual incomes. In some cases, the conclusion of an inverted U-shape relationship between growth and inequality was reinforced, as in Rauch (1993), for instance, while in other cases, the inverted U-shape conclusion was undermined – see for instance the demand-based models by Taylor and Bacha (1979), de Janvry and Sadoulet (1983) or Bourguignon (1990).

It is not clear that the inverted-U shaped relationship between inequality and the level of development is an important issue by itself. It may have been important in the debate of the early 1970s on whether development efforts had to concentrate solely on growth as opposed to growth and distribution.[9] What seems more important today is the recognition that at all stages of the development process economic growth has indeed the capacity of modifying social structures, and in particular the hierarchy of relative incomes among individuals or socio-economic groups, through its natural sectoral bias. The way in which this bias is actually translated into more or less inequality is likely to be country-specific, but this strand of the literature suggests that it is not justified a priori to start from the postulate that economic growth is neutral in the sense that it affects everybody's living standard in the same proportion.

It is interesting that this emphasis on the sectoral bias of growth as the source of social changes is still present in the recent literature on inequality. Indeed, several authors see the appearance of a new branch in the Kuznets curve in the surge of wage and income inequality observed in the US economy in the late 1970s and in the 1980s – see for instance List and Gallet (1999). This evolution is often interpreted as the consequence of technical progress and/or international trade. A new sectoral specialization is appearing and social structures are affected by a progressive transition of the whole labor force towards the most modern sectors of the economy, in a process reminiscent of the industrialization process behind Kuznets' original argument.[10]

(b) General Equilibrium Models of the Distributional Effects of Growth

The sectoral shift view at the distributional consequences of growth refers to a 'fix-price' view of economic development where population movements across sectors or socioeconomic groups respond to some disequilibrium. This disequilibrium may itself be caused by economic growth, but that relationship and the growth process itself are not explicitly considered. An alternative approach to the distributional effects of economic

9 Chenery et al. (1974) provides a good summary of the various elements of this debate. A detailed account of the recent history of economic thought in this area is offered by Arndt (1987).
10 For empirical evidence from the US, see Katz and Autor (1999) and Baumol, Blinder and Wolff (2003).

growth consists of considering changes in factor prices that may take place along the growth path, together with the factor accumulation behavior that is causing growth. Such an analysis is equivalent to linking the micro-economic analysis of distribution with standard macro-economic theories of growth and the functional distribution of national income. Sectoral differences and disequilibria which were prominent in the preceding approach are now ignored because it is now implicitly assumed that factor and good markets are permanently in equilibrium. The theoretical framework thus is that of dynamic general equilibrium rather than that of temporary fix-price partial equilibrium.

Numerous dynamic general equilibrium models have been proposed to analyze the relationship between economic growth and inequality, many of them inspired by the recent revival of the growth literature. No attempt will be made here to give an exhaustive summary of that literature, which in effect tends to concentrate on the way inequality affects economic growth rather than the opposite.[11] Instead, this section looks at the other face of the coin, namely what is to be learnt from this literature about the distributional consequences of growth.

Following the pioneering paper by Stiglitz (1969), assume that the income, y_i, of agent i comes on the one hand from labor and on the other hand from the return on his/her wealth, k_i. To simplify, assume that all agents have the same labor productivity and supply one unit of labor. Thus their labor income is uniform and equal to the wage rate, w. These assumptions are more general than they look at first sight if k_i incorporates both physical (or financial) and human capital. The uniformity of labor income thus is an assumption that tries to represent the fact that (raw) labor income is in general more equally distributed than physical or human capital. Generalizing the following argument to the case of some exogenous distribution of labor productivities does not raise major difficulty. Denoting the rate of return to capital by r leads to the following definition of individual income:

[11] For a survey of this literature, see Aghion et al (1999), Benabou (1996) or Bertola (2001), as well other chapters in this volume.

$$y_i = w + r.k_i \qquad (2)$$

Expression (2) shows a first way through which growth may affect the distribution of income among individuals. By modifying the relative rewards of labor and capital, growth modifies relative individual incomes. For a given distribution of wealth, the distribution becomes more equal when the relative reward of labor rises.

As growth proceeds through the accumulation of individual wealth, there is a second way by which it may modify the distribution of income and wealth. A simple general assumption is that saving or investment by agent *i*, S_i, is a linear function of the various sources of income :

$$S_i = \alpha.r.k_i + \beta.w - \gamma \qquad (3)$$

where α is the marginal propensity to save out of capital income, β out of labor income and γ stands for the effect on savings of the existence of some minimum consumption level. Assuming in addition a depreciation of capital at rate δ, for all agents i, leads to the following accumulation behavior :

$$\dot{k}_i = (\alpha r - \delta).k_i + \beta w - \gamma \qquad (4)$$

or in growth rates :

$$\dot{k}_i / k_i = (\alpha r - \delta) + (\beta w - \gamma) / k_i \qquad (5)$$

where the notation $\dot{}$ refers to infinitesimal time change.

It may appear that the preceding specification only allows for the representation of the evolution of the distribution of income and assets among individuals and does not permit analyzing social structures as described by the composition and relative income of socio-economic groups. This is not the case, however. It is sufficient to define socio-economic

groups by some particular endowments of assets for (1) to describe the evolution of the relative incomes between, for instance, unskilled workers and skilled workers, or between workers and 'capitalists'. In effect, what matters here is the return to the various types of assets that are used to define socio-economic groups. The evolution of these rates of return defines the evolution of between group inequality. Likewise, equations (4)-(5) implicitly define the dynamics of group composition. For instance, some unskilled workers—with k initially equal to zero—acquire some positive human capital (k) and become skilled workers, whereas some skilled workers may become 'capitalists'. Thus, it should be clear that all that follows applies as well to a description of the effects of growth on inequality among individuals as well among socio-economic groups on social structures.

An obvious implication of the linearity of (4) is that the distribution of wealth and income *does not affect* the aggregate growth path of the economy and therefore the evolution of the factor prices, w and r, that comes with it. Yet, it can be seen in (5) that another implication of that saving function is that *growth generally induces a change in* wealth or income inequality. In the general case, inequality decreases or increases with growth depending on whether savings out of wages, βw, cover the dissaving due to minimum consumption, γ, or not.[12] Since the wage rate is expected to increase with growth, inequality may increase with growth in a poor economy but this evolution may revert itself when the economy has reached a certain level of affluence, in a process that is consistent with the Kuznets hypothesis.

This is an extremely simplified model. At the same time, it incorporates enough flexibility to analyze several interesting issues. To do so, it is helpful to derive from the preceding equation the time behavior of relative incomes. It is easily shown that the evolution of the relative income of two agents i and j is given by :

[12] Results would not be qualitatively different if it were assumed that savings cannot be negative but are nil if income is below minimum consumption – see Stiglitz (1969) and Bourguignon (1981).

$$\frac{\dot{y}_i}{y_i} - \frac{\dot{y}_j}{y_j} = \left(\frac{1}{y_j} - \frac{1}{y_i} \right).\{w.(\dot{r}/r - \dot{w}/w) + w.(\alpha r - \delta) - r.(\beta w - \gamma)\} \qquad (6)$$

This expression shows that distributional changes along a growth path have various sources. The first term in the curly bracket on the RHS corresponds to the changes in factor prices resulting from growth, whereas the last two terms correspond to capital and non-capital sources of savings, respectively. Sources for distributional changes are thus richer than with the sectoral shift approach. As mentioned above, the accumulation component may be compared, to some extent, to population shifts in the sectoral model. Differential accumulation behavior in the present representation of growth is equivalent to individuals moving from one socio-economic group to another. But, of course, it also corresponds to possible changes in 'within-sector' inequality parameters (V_i). The factor price effect corresponds to a change in the income differential across groups, that is the ratio y_1/y_2 in the sectoral shift model.

Expression (6) readily shows what evolution in factor prices and what kind of saving behavior may be responsible for increasing or decreasing income disparities among individuals or socio-economic groups along the growth path. According to the factor price effect, any increase in the reward to capital, relative to labor, increases inequality by lowering the relative income of 'pure' workers. The same is true of a high propensity to save out of capital income. On the contrary, a high propensity to save out of labor incomes contributes to more equality in the economy, at least after a wage threshold has been passed.

Of course, accumulation behavior and factor price changes cannot be considered as independent of each other. In this respect, it is interesting to consider some particular cases of the preceding general model. A first simple case is when agents save only out of their capital income ($\beta = 0$) and there is no minimum consumption requirement ($\gamma = 0$). It is well known that such a saving behavior can be obtained as the implication of a simple

life cycle consumption allocation model.[13] As can be seen in (5), the rate of growth of individual wealth is then the same for all agents with positive initial wealth, which implies that the distribution of wealth remains constant over time, maintaining the features inherited from history. Note that this does not necessarily mean that the distribution of income will remain constant since relative factor prices and factor shares in individual incomes may change along the growth path. Yet, in the standard neo-classical framework with unit elasticity of substitution between capital and labor, it is easily shown that the distribution of relative incomes remains constant over time, precisely because the factor shares of total and individual incomes are constant. [14]

A related particular case, very much emphasized in the recent literature, arises when $\beta = \gamma = 0$, the rate of return to capital is constant and the wage rate grows at the same rate as individual and aggregate wealth. This would correspond to an economy where output is proportional to capital and is divided in constant proportion between labor and capital. The implicit growth model behind this description could be a version of the Harrod model or the 'aK' endogenous growth model proposed by Frankel (1962), extended later by Romer (1986) and others. The implications of these two particular cases are worth to be stressed. They indeed provide an interesting benchmark where *economic growth is essentially distribution neutral*, even after taking into account both the process of wealth accumulation and its effects on good and factor markets. However, it can be seen that this result is not so much due to the assumptions made on the production side of the economy - i.e. constant or declining returns to scale – as to the assumption that savings arise only out of capital income.

Another interesting particular case is the one originally explored by Stiglitz (1969). Assume $\alpha=\beta$ and that factor prices are determined by the marginal products of an

[13] In an infinite horizon model it is necessary to assume first that the inter-temporal elasticity of substitution (σ) is constant. If r is the rate of interest and ρ the utility discount rate, this leads to a constant rate of optimal consumption $g = \sigma.(r-\rho)$. It is sufficient to assume that non-capital income grows at the same rate g, as will be the case at the steady state in an economy with constant returns to scale, to obtain that savings are proportional to existing wealth – see Bertola (1993).

[14] The unit elasticity of substitution ensures that aggregate shares of capital and labor in total income are constant. As all individuals accumulate capital at the same rate, individual shares of total capital remain constant. Together, this implies that both individual capital and labor incomes grow at the same rate as the whole economy.

increasing and concave aggregate production function. Then, the aggregate economy behaves as in the well-known Solow model (1956) with the distribution of income following a Kuznets curve if the wage rate is initially low enough. More interestingly, and somewhat paradoxically, it can be shown that the distribution of income and wealth tends asymptotically towards full equality if the steady-state wage/profit rate ratio is large enough.[15] This result is considerably weakened if labor productivity and labor incomes are assumed to be heterogeneous across individuals. The distribution of both wealth and total income may then be shown to converge asymptotically to the distribution of labor productivities.

(c) Non-linear Savings Behavior

The preceding results are all based on the assumption that the wealth accumulation process underlying growth is linear with respect to individual wealth. This assumption is debatable. There are various reasons why savings may be thought to be a non-linear function of income or wealth, even when saving behavior is strictly assumed to result from inter-temporal optimization. Liquidity constraints and/or credit market imperfections are the most obvious factors that may explain such non-linearities. For instance, credit rationing may imply that zero is a lower bound for the savings of somebody with zero wealth. Combining this feature and the preceding linear model leads to savings being defined by :

$$S_i = Inf\{0, \alpha.r.k_i + \beta.w - \gamma\}$$

[15] Namely, w/r>γ/δ at the steady state. The proof of the convergence towards equality is simple. (5) may be rewritten as $\dot{k}_i = \left[\alpha(r\bar{k} + w) - \delta.\bar{k} - \gamma\right] + \left[(\alpha r - \delta)(k_i - \bar{k})\right]$ where $\bar{k}$ is the mean wealth in the economy. At the steady state of the economy, the first square bracket on the RHS is nil and the first term in the second square bracket is negative. It follows that individual wealth necessarily converges towards the mean wealth, $\bar{k}$. In comparison with the general model, this proof shows that the equal distribution asymptotic result in this particular case is due to : a) the equal marginal propensity to save out of capital and labor income; b) decreasing marginal returns to production factors.

In effect, this apparently small modification is sufficient to drastically alter the conclusions obtained previously. First, aggregating individual accumulation behavior leads to a change in aggregate wealth that depends on the distribution of wealth. Thus, the distribution of wealth affects the aggregate growth path of the economy, which was not the case before. Second, some of the conclusions obtained previously on the evolution of the distribution of income and wealth do not hold anymore. In particular, the result that the distribution of income tends towards equality in the Solow-Stiglitz model is not anymore granted. Depending on the initial distribution of wealth, inequality may well be non-decreasing throughout the whole growth path of the economy. Analogous conclusions may be obtained with more general non-linear specifications of the saving function.[16]

A way of justifying non-linear saving functions is to account for the fact that the rate of return to capital in the original (linear) model above may be heterogeneous across agents with different levels of wealth or income. Credit market imperfections associated with the existence of some indivisible investment project with an exogenous rate of return are sufficient to generate such a result. For moral hazard or adverse selection reasons those individuals who have to borrow to undertake this project face a borrowing rate of interest above rates served on conventional savings – and possibly decreasing with the amount borrowed, or equivalently their wealth.

Under the preceding assumptions, the original individual accumulation equation writes then :

$$\dot{k}_i = \alpha[r(k, k_i).k - \delta.]k_i + \beta w - \gamma \qquad (7)$$

[16] Of course, the inegalitarian steady state result may be obtained for any saving function that is convex with respect to income or wealth. Schlicht (1975) and Bourguignon (1981) offer a general treatment of convex saving functions within a Solow-Stiglitz framework, without analyzing the reasons for convexity. Moav (2002) reaches the same conclusions assuming intertemporally maximizing agents with convex bequest functions.

where r( ) is a function of k_i that has the shape of an inverted U – poor people are credit rationed and only people in some intermediate wealth range are borrowers facing an implicitly higher rate of return on their wealth.

A significant proportion of the recent literature on the effects of the wealth distribution on economic growth is implicitly based on credit market imperfections and an accumulation equation of type (7). This is true in particular of the seminal papers by Galor and Zeira (1993), Banerjee and Newman (1993), Aghion and Bolton (1996) and Piketty (1997). These models also have implications for the distributional consequences of growth. As the rest of the literature, however, they suggest that all types of evolution are possible, from continuously increasing or decreasing inequality to Kuznets-curve-like movements.

d) The role of technical progress

The preceding is based on a view of economic growth being essentially driven by factor accumulation. Growth modifies the distribution of income and wealth among individuals or across socio-economic groups essentially because individuals or groups do not accumulate at the same rate and because factor accumulation may cause changes in the remuneration of the productive factors owned by individuals. But, of course, another engine of growth is technical progress, which may itself modify both the relative remunerations of productive factors and factor accumulation behavior.

If technical progress was neutral, affecting the remuneration of all factors in the same way, then nothing would have to be changed in the preceding argument. An issue arises when technical progress is 'biased' in the sense that it favors one factor more than others. A case that has received very much attention in the recent literature is that of the 'skill-biased' technical change, which is a shift in technology that increases the demand for skilled labor.[17]

[17] See Acemoglu (2002), Katz and Autor (1999)

From a theoretical point of view, one may analyze the effect of skill-biased technical change on social structures as resulting simply from a change in the return to the human capital component of individual wealth in the preceding framework. The increased demand for skilled labor increases the return to skill and, other things being equal, increases the income differential between skilled and unskilled workers. Of course, an increase in the return to skill is likely to cause an acceleration in the human capital accumulation, reducing progressively the differential between skilled and unskilled workers and, at the same time, shifting workers towards high skill socio-economic groups. The effect of this 'race' between technical change and education or training was first analyzed by Tinbergen (1975). Since then, it received very much attention.[18] In effect, the implications of this race for inequality are a priori ambiguous since it depends on the speed at which the demand for skill labor increases following some technological innovation and the speed at which skill accumulation responds on the supply side of the labor market. If growth is seen as successive waves of skill-biased technical innovations, it may thus be accompanied by long-run fluctuations in earning differentials across skill groups of workers. These fluctuations may also be influenced by the fact that the bias of technical change may itself be affected by the skill differential and the relative availability of skilled workers, as in Acemoglu (2002).

The dynamics of the earning differential is not necessarily as simple as the preceding supply-demand argument would imply. For instance, Aghion et al. (1999) and Aghion (2002) develop an original model of the diffusion of an innovation in General Purpose Technology where skill-biased technical change contributes to a continuous increase in the skill wage differential, after a preliminary period where the differential remains constant. The skill gap starts increasing when all skilled workers have been absorbed by firms which have adopted the new technology and the gap keeps increasing as long as other firms seek to adopt the new technology too. All firms eventually adopt the new technology and all workers acquire the new skill. In effect, this process combines both a sectoral shift of the type analyzed above and a general equilibrium price effect.

[18] For recent formalizations of this argument see Eicher (1996), Galor and Tsiddon (1997). See also the survey by Aghion et al. (1999).

Although most of the effect of skill-biased technical change is expected to take place between skill groups, some authors insist that it may also affect inequality within groups. Any innovation requires adaptation of the first workers who are confronting it, and this adaptation is easier for workers with some specific ability on top of the skill required to perform the new task. The increased remuneration of that specific ability contributes to increasing the degree of inequality among workers, an inequality that may persist over time if the adaptation to the new technology has created a new type of human capital among the workers who were first exposed to the technical innovation (see Violante, 1996, Rubinstein and Tsiddon, 1998, Aghion, Howitt and Violante, 2002).

<Leave some space>

To conclude this section on the theoretical mechanisms through which growth affects social structures, it must be emphasized that, as mentioned in the case of technical progress, sectoral shift and general equilibrium mechanisms are not necessarily mutually exclusive. It was already seen above how the individual asset accumulation equations (4)-(5) could actually represent the shift of individuals across socio-economic groups defined by their factor endowments. More explicitly, however, recent theoretical models built on the imperfect credit market mechanism actually combine the factor market and the sectoral shift approach. An interesting aspect of the model proposed by Banerjee and Newman (1993), for instance, is that the distribution of wealth in the economy practically determines economic agents' kind of occupation and the sector where they operate. People with little wealth are pure wage workers, employed either in the formal or the informal sectors. People with a higher initial level of wealth engage in small businesses and determine the size of the 'informal' sector, whereas richer people are the owners, managers and top employees in the formal sector. The change in the wealth distribution that takes place, together with growth, along time has thus the effect of changing the distribution of occupations in the economy and the relative size of the formal and informal sectors. To some extent, this particular dynamic general equilibrium model

provides a kind of formalization of the intersectoral shifts emphasized in the Kuznets tradition of the analysis of the effects of growth on inequality.

Overall, the short preceding review shows that considerable efforts have been devoted to identifying and understanding the mechanisms through which growth affects social structures and inequality. As far as social structures are concerned, the analysis points to the evolution of the employment structure of the population away from lowest productivity sectors and occupations. An obvious social consequence of growth thus is to reduce the share of the population living in traditional agriculture in a first stage, in informal non-agricultural activities in a second stage, in 'low-tech' manufacturing activities in a third stage, etc. At the same time, the accumulation of human capital implies that growth comes with a continuous reduction in the share of the population with no or low education.[19] On the other hand, the economic analysis of growth is largely inconclusive concerning other aspects of social structures and the distribution of wealth or income.

In this respect, three sources of ambiguity must be stressed. First, theory is necessarily silent on aspects of social structures that do not appear as central in economic growth mechanisms. That the population shifts sector and occupation in a well defined direction is one thing. Whether this movement is uniform across population subgroups defined by gender or ethnicity is another thing—about which economic growth theory has little to say . This is essentially because factor endowments put forward by growth theory do not incorporate this dimension of social differentiation. Second, whether the shift of the population from one sector or socio-economic group to another is accompanied by an increase in income differentials among those sectors or groups is unclear. For instance, the share of educated people in the population is increasing with growth, but that evolution is in theory consistent with constant, increasing or decreasing income differentials between educational levels. Third, if economic theory permits to identify the

[19] The preceding argument is straight economics. But the increase in schooling has also a demographic explanation. The demand for children—and thus family size and human capital levels—changes with economic growth. One tends to observe less children of better 'quality' at higher levels of income per capita. For surveys of this literature, see the two volumes of Rosenzweig and Stark (1997) and Rosenzweig (1990).

various channels through which growth may affect inequality among individuals, the sum total of these effects is ambiguous: no change, equalizing or unequalizing evolution throughout the whole growth process, or the inverted U-shape put forward by Kuznets. This conclusion holds whatever the analytical framework being used, whether theoretical models belong to the sectoral shift fix-price or to dynamic general equilibrium modeling tradition.

3. THE EFFECT OF ECONOMIC GROWTH ON SOCIAL STRUCTURES : EMPIRICAL EVIDENCE

The preceding analysis suggested various channels through which growth is affecting social structures: a) by shifting individuals from one sector or socio-economic group to another; b) by modifying income differentials across sectors and socio-economic groups; c) by modifying income and welfare disparities across individuals. Of course, these three channels are not independent. In particular, it must be clear that sectoral shifts and inequality changes between socio-economic groups have a direct impact on inequality among individuals.

This section briefly reviews the empirical evidence available on these three channels. Both on a cross-country and case study basis, the sectoral shift effect of growth turns out to be the fundamental way through which economic growth affects social structures. Cross-country analysis is relatively less conclusive both for differential effects of growth across socio-economic groups or among individuals. However, this certainly does not mean that growth has no impact on social structures outside the sectoral shift component. Rather, case studies suggest that this effect is relatively more complex and most likely strongly country-specific.

(a) The Sectoral Shift Effect of Growth on Social Structures

Two rows in table 1 illustrate the power of growth related sectoral shift to influence on social structures. Urbanization and all the phenomena that it entails is a the first example. The structural explanation behind it is clear. It has to do with the falling share of agriculture – or, better said, of low-productivity traditional agriculture - throughout the growth process. Although only a reduced form model appears in table 1, coefficients shown there are strongly significant and it would be relatively easy to devise a structural model focusing more on the mechanisms behind the urbanization process with equally strong statistical significance.

A second sectoral shift effect of growth shown in table 1 concerns education. Here again, the positive correlation with development levels is very strong. It is true that the coefficient obtained with decadal differences is not statistically significant, but this might have to do with the extremely long lags with which changes in schooling behavior, possibly generated by economic growth, spread to the whole population. [20] It is also true that focusing on literacy rates yields a perspective on education and skills that may seem too narrow. However, it is unlikely that considering the proportion of 'skilled workers' in the labor force - assuming that some uniform definition of skills is available across countries- or the proportion of people with secondary education would yield very different qualitative results.[21] In effect, regressing the average number of years of schooling of individuals in the labor force on GDP per capita in table 1 yields results qualitatively similar to the regression on literacy rate.

One could undoubtedly multiply regressions showing strong structural effects of economic growth through which changes in social structures are likely to occur. For instance, one could focus on the weight of the manufacturing or the service sector instead of focusing on agriculture, or one could focus on the relative weights of low- and high-tech industries or enterprises. Interestingly, this kind of approach to growth which has

[20] From this point of view, empirical evidence may seem more pertinent . See Clemens (2004), Lindert (2003) and Krueger and Lindahl (2001).

[21] For a more detailed analysis of the way in which the structure of the population by educational level changes with the level of development, see Thomas, Wang and Fan (2000).

been prominent in the 1970s, as a continuation of Kuznets research program on 'modern economic growth' and under the impulsion of Chenery and associates [22] is presently weakening. There are various reasons for this neglect, in particular for developed countries as will be seen below. Yet, it would be wrong to conclude from this relative lack of interest that sectoral shift phenomena have disappeared from the research agenda when analyzing the social effects of growth. Indeed, the whole recent literature on the skill bias in the sources of growth and, in particular, technical progress or international trade (see, for instance, the survey by Katz and Autor, 1999, and Baldwin and Cain, 2000) may be considered as an updated sectoral shift argument in the analysis of growth and its effects.[23]

Changes in female labor force participation may also be considered as a sectoral shift phenomenon. But it may also be considered as deriving from a change in behavior itself caused by economic growth. According to the sectoral shift logic, women would be moving from being 'inactive' or more exactly specialized in low-productivity domestic production to market employment at a higher level of productivity. According to the behavioral interpretation, the role of women would have been changing in a way concomitant with growth but under forces of a different nature. For instance, some authors see an explanation of the increased female labor force participation in most developed countries after WWII as resulting from the excess demand in the labor-market that developed during the war and that had to be filled, a phenomenon that produced a durable change in behavior. Others would insist that the drop of fertility partly due to the diffusion of contraceptive means in the last three decades or so was the reason behind women's increased labor force participation.[24]

---

[22] See Chenery et al. (1974) and Chenery and Syrquin (1975). For a more recent statement and a comparison with the contemporaneous growth literature, see also Syrquin (1994).

[23] It is true that much of that literature is concerned with changes in skilled/unskilled wage differential, an issue which we take up in the next subsection, but the basic argument—which goes back to Tinbergen (1975)—is that the evolution of technology which is behind economic growth requires more educated labor.

[24] The first hypothesis is critically reviewed by Goldin (1991). For the second, see for example Birdsall and Chester (1987), Asbell (1995), Goldin and Katz (2002).

The preceding phenomena may well provide a partial explanation for the fast increase in female labor force participation in the developed world over the last 60 years or so. However, regressions in table 1 reveal a rather strong association between participation and economic growth that is not due to cross-country differences, whereas both the fertility and the WWII argument would suggest that those cross-country differences should dominate. Pure cross-country regressions in the first columns of table 1 yield unsignificant results or wrongly signed coefficients when participation is specified as a linear function of GDP per capita. On the contrary, strongly significant results are obtained when within country time behavior of participation and GDP per capita is taken into account. The simple structural argument that increased female labor force participation may correspond to a shift from low to higher productivity occupations is not undermined by the data. Interestingly enough, the same difference between cross-country and within country estimates appears when participation is regressed on a quadratic form of GDP per capita. A U-shape relationship is obtained when not controlling for country fixed effects, whereas a monotonic relationship holds in the opposite case. [25]

For the sectoral shift effect of growth on social structures to be of relevance, it is necessary that it takes place between sectors or socio-economic groups with sufficient initial important differences in terms of welfare level. This is certainly the case for the shift across skill (or education) levels, between traditional agriculture and the modern sector of the economy in developing countries or between inactivity and market work for women. Things are less clear in the case of sectoral shifts in developed countries when markets function smoothly and tend to equalize returns of human capital across occupations or sectors. Most likely, this is the reason why this dimension is somewhat neglected in the recent growth literature in developed and emerging countries. The attention there tends to concentrate on differences in the evolution of returns to assets and in their accumulation rather than their sectoral allocation.

[25] On the U-shaped female labor force participation function see Durand (1975) or Goldin (1995). The decreasing part of the curve is explained by the decline in female unskilled employment opportunities due to the contraction of the agricultural sector, whereas the increasing part would be due to rising education. See, for instance, Clark, York and Anker (2003). In table 1, an inverted U-shape is obtained when controlling for fixed effects, but the top of the curve occurs at income levels much above what is observed in the sample.

The previous remark illustrates possible differences across countries that are hidden by the cross-country work that has been referred to. That differences do exist between developed and developing countries in terms of the social consequences of sectoral shifts of population due to growth is obvious. In table 1, these differences often are accounted for through the Cox transformation on the dependent variable. But other country differences might exist so that one would ideally like to estimate sectoral shift equations using national time series rather than a cross-section of countries. To some extent, this country specificity is what Chenery and his associates were after when they tried to identify 'patterns of development' among developing countries. Unfortunately, due to lack of adequate data, they most often had to rely on calibrated structural models rather than time series structural econometrics. The situation has changed little since then.

(b) Effect of Growth on Inequality between Socio-economic Groups

As mentioned above, the sectoral shift effect of economic growth is important to explain social structures inasmuch as it is accompanied by a persistent differential between socio-economic groups or sectors, in terms of current of permanent income or welfare level. At the same time, it may be envisaged that economic growth contributes to a deepening, or on the contrary, to a weakening of this social differentiation. In effect, those two evolutions are certainly not independent. Sectoral shifts need income differentials to develop and, in turn, they produce changes in these differentials. This subsection focuses on the potential effects of growth on earnings or income differentials across socio-economic groups. Given the dualism with the sectoral shift argument, we adopt the same presentation as in the previous subsection and consider in turn income differentials between sectors - essentially agriculture and the rest of the economy, between skills or educational levels, and between genders.

Sectoral Income Differentials

The regressions in table 1 illustrate the fact that sectoral productivity - and presumably income - differentials tend to diminish with economic growth. Thus, growth contributes to harmonize social structures across sectors. The process behind this is clear and has indeed very much to do with the sectoral shift process. As growth proceeds and high productivity activities arise, people tend to leave low-productivity occupations predominantly located in traditional agriculture. But, this migration process contributes to increasing productivity and income in the sector of origin - and possibly to lower income growth in the sector of destination. The 'dualism' of the economy, emphasized by early development economists, tends to diminish with growth. It eventually vanishes when the economy is more mature and market mechanisms ensure the equalization of productivity and earning rates across sectors.

At early stages of development, the preceding process is undoubtedly a powerful source of changes in social structures. At later stages, the emphasis on the agricultural sector is probably ill-placed. A comparison between informal and formal (non-agricultural) sectors would be more appropriate. If comparable data were available across countries on this formal/informal distinction, they would probably show a similar phenomenon, that is a narrowing of productivities and incomes across sectors as the informal sector loses weight. At some point, however, the issue becomes essentially that of the functioning of the labor market. The difficulty then is to identify whether earning differentials are due to some segmentation of the labor market or correspond to the self-selection of individuals across jobs with different characteristics and productivity levels.[26] Undoubtedly, these differences raise important social issues regarding the social status differential of workers linked to workers' social status. But they are of a nature different from the social transformations taking place at earlier stages of development, and it is not clear whether they may be unambiguously associated with growth.

Effect of Education on Earnings

[26] For references, see Ashenfelter and Card (1999). See also Card (1999) which reviews the methodological problems involved in estimating the causal effect of education on earnings.

In a competitive factor market environment, it was argued above that differences in productive asset bundles owned by individuals would be a better indicator of social differentiation than the sector of occupation. Education was thus seen as an important dimension of social differentiation. In this context, the sectoral shift analysis coincides with the change in the distribution of the population, or possibly the labor force, in terms of educational levels. It is now time to examine whether earning differentials across educational groups - often assimilated to skill groups - tend to change in a systematic fashion with economic growth.

There is a huge literature on earning differentials by educational or skill levels and their evolution over time. This is not the place to summarize it.[27] There is considerably less literature on comparing differentials across countries. To our knowledge, the main contributor in this area is Psacharopoulos who devoted very much effort to the collection of rates of return to education derived from the estimation of Mincerian earning equations based on labor force or household surveys around the world - see in particular Psacharopoulos (1994) and Psacharopoulos and Patrinos (2002). Putting together Psacharopoulos' findings and the lessons from country studies on the evolution of earning differentials leads to some interesting and somewhat paradoxical conclusion. Namely, cross-country comparisons suggest that there is a strong long-run tendency for earning differentials across educational levels to fall, whereas country studies show a very high level of medium-run variability without clear trend.

<Table 2 around here>

Table 2 shows mean earning differentials across schooling levels for country groups defined by GDP per capita. These groups cover a total of 98 developed and developing countries for which Mincerian earning equations were available during the period extending from 1970 to 1996, the most recent estimate being used (usually from the late 1980s or early 1990s). All individual earners are supposed to be covered by the data, whether they are wage earners or self-employed. The striking fact is that earning differentials tend to decline with the level of average income above primary, whether the differential is taken between secondary and primary, or between tertiary and secondary.

[27] See Katz and Autor (1999)

The last column of the table shows the average rate of return by year of schooling for the same data sets. They, also, fall with the level of GDP per capita.

This empirical regularity—which nevertheless hides very much variability across countries—is consistent with a simple competitive story of the labor market. As the accumulation of human capital proceeds, the return to skill tends to fall. If cross-country differences are taken to represent the effect of long-run growth, then the idea behind this story is essentially that the demand for skilled labor tends to grow at a slower pace than supply, thus reducing gaps between educational groups. Interestingly enough, however, time series analyses for countries where successive estimates of the Mincerian equation are available tell a different story. Restricting now the analysis to the average rate of return by year of schooling, it can be seen in figure 1 that practically all evolutions are possible over period extending from 8 to 23 years. Near constancy as in the case of Brazil to steady increase as in the US between 1976 and 1990, a phenomenon extensively studied in the literature, to steady decline as in the case of Netherlands, to erratic behavior as in Sweden.

<Figure 1 around here>

There are various reasons behind this paradoxical difference between cross-sectional and time series evidence. First, time variations in returns per year of schooling may have different origins depending on the level of development. They may originate more across primary and secondary in developing countries and more between secondary and tertiary or even within tertiary in developed countries.[28] This is because practically everyone has 8 years of schooling or more in developed countries, whereas most low-income and middle-income countries are far from that goal. Second, there may be problems of comparability of samples across countries at different levels of development. In particular, most earning equations are estimated on samples of urban workers. This group of workers represents a high percentage of the labor force in high-income countries, much less in others. Third, the period on which time series are observed may not be long enough to be fully consistent with cross-country differences in development levels.

[28] This is a natural consequence of the fact that schooling until middle secondary has been compulsory in developed countries for quite a long time.

Finally, there are reasons to expect some time specificity in the evolution of the rate of return to schooling and skill differentials in general. As alluded to before, they are related to Tinbergen's (1975) idea of the race between technology and education driving the evolution of skill differentials in earnings. For more or less continuous progress achieved on the educational side, it is sufficient to imagine fluctuations or trends in a definite direction in skill-biased technical progress - or possibly in trade policies - to generate time series behavior of the type shown in figure 1. Alternatively, one may also think of technical progress and its diffusion in the economy following a continuous – not necessarily linear – trend and the educational response intervening with a different time profile to generate various patterns in the evolution of the rate of return to schooling.[29]

For all preceding reasons, it is not contradictory to infer from existing evidence that, indeed, a narrowing of skill differentials and a fall in the rate of return to education is accompanying growth in the long-run, at least in the early stages of development. At the same time, however, this trend may be hidden for long or short periods by accelerations of skill-biased technical progress, major changes in the international trade environment of a country, and velocity of the supply response to changes in skill differentials.

Gender Earnings Differentials

The examination of evidence on gender gaps in individual earnings leads to conclusions opposite to the preceding ones. There, one observes a rather clear downward trend in most countries, although information is scarce for developing countries. By contrast, cross-country comparison points to no systematic differences between countries ranked by development level.

There are relatively few cross-country comparisons of male-female earnings differentials in the literature except for developed countries. Terell (1992) collected estimates on

[29] See for instance Aghion (2002). See also the chapter by Jorgenson in this volume.

male-female earning differentials from the ILO Yearbook of Labour Statistics. She found very much cross-country variability, and more importantly as much variability within low and middle-income countries as within high-income countries. For instance the female-male earnings ratio varied from .60 to .85 in developed countries, and from .50 to .90 in developing countries. Of course, the problem is that the samples on which these ratios are evaluated, typically salaried urban workers, differ quite substantially across countries. Thus, the comparison may be of little relevance between a country where 80% of both men and women are in that group and countries where only 20% of men and still a lower proportion of women qualify. This is what explains the paucity of cross-country comparisons encompassing the whole world. Better comparisons can be performed using indirect indicators of earnings. For instance, the regression on male-female difference in literacy rates in table 1 shows a steady decline with the development level both across-countries and across 10-year periods. However, literacy or schooling differences is only one among many determinants of earnings differences. Not much is known about gender differences in these other dimensions, nor about their impact on earnings differentials. [30]

In contrast to cross-country, time series of female-male earning ratios available in developed countries show an unambiguous increase over the last few decades. In the US, for instance, this ratio increased from .56 to .72 between the late 1960s and the late 1990s - see Welch (2000). Comparable increases are observed in European countries .[31] Only a few data points are available for middle-income countries but they reflect a parallel evolution. Between the early 1980s and the mid 1990s, the female-male ratio among wage earners increased from .74 to .80 in Taiwan, from .62 to .72 in Colombia, and from .52 to .61 in Brazil.[32]

[30] The life expectancy regression in table 1 also shows an evolution that seems relatively more favorable to women along the growth process in pure cross-section. Yet, the sign of the coefficient is reversed when fixed effects are taken into account and the coefficient is unsignificant in decadal differences.. More detailed characteristics are analyzed in the World Bank (2001).
[31] For a thorough and exhaustive discussion of these issues, see for instance Gunderson (1989).
[32] The latter figures are taken from Bourguignon, Ferreira and Lustig (2004)

The problem is to know whether such an evolution should be related to economic growth or not. This requires distinguishing between two sources of change in female-male earnings ratios, one associated with the differential evolution of the characteristics of female and male workers, and the other with changes in the remuneration of these characteristics - i.e. the well-known Oaxaca-Blinder decomposition. In effect, it so happened that the increase in female wage employment concomitant with the previous evolution contributed overall to a lowering of the average skill of employed women in many instances, an effect contributing to a widening rather than a narrowing of the gap. Thus, the observed narrowing must prominently be due to changes in remunerations rather than in the differential characteristics of male and female workers.

The issue then arises of the source of changes in remunerations. Are they related to specific observed characteristics or more diffuse in the whole labor force? In the latter case, it would then be tempting to associate the decline in the female-male ratio to non-economic phenomena like the evolution of social norms about gender differences, to regulation in the labor markets - minimum wage for instance - or, in some countries, to some kind of affirmative action policies. All these phenomena probably bear some responsibility for the narrowing of the male-female gap, but there also is indirect evidence that it may be due to a change in the remuneration rate of some specific characteristics that distinguishes male and female work. This indirect evidence is the positive correlation that has been noted between the female-male earnings ratio and the degree of inequality of male earnings, both in time series -see Blau and Kahn (1997) and Welch (2000) - and in a cross-section of developed countries - see Blau and Kahn (2003). A possible interpretation of that correlation is that changes in male earnings inequality reflect changes in the remuneration rate of earnings determinant, the relative intensity of which happens to differ much between male and female labor. As observed worker characteristics like education or experience do not appear to have played this role, this explanation must rely on unobserved earnings determinants. For instance, Welch (2000) refers to "brains relative to brawn", with the idea that the relative remuneration of brains in the labor market increases with technical progress, and in effect with economic growth, whereas woman labor is typically more intensive in that factor.

Such an interpretation of the narrowing female-male earnings gap in high-income - and possibly some middle-income countries - is attractive. Yet, it remains to be tested more carefully, an uneasy task given that this hypothesis is essentially based on unobserved labor characteristics. At this stage, it is thus difficult to conclude whether it is indeed economic growth that so strongly influences this fundamental dimension of social differentiation. At the same time, it is worth stressing that the preceding issues probably arise only beyond some development level where the labor market is sufficiently unified and competitive. The thinness of the modern labor market and the low weight of women in that market may explain why low income countries are not really comparable to others in terms of gender earnings differentials.

(c) Effects of Growth on Inequality among Individuals

The preceding subsections looked into the effect of growth on the structure of the population by sector of activity and by socio-economic groups defined by some observed characteristics, and on income differentials between them. It is now time to consider the possible effect of growth on the overall distribution of income among all individuals in the population. Such a perspective goes beyond the preceding points of view in that it adds to the analysis the possible effect of growth on unobserved individual characteristics through changes in disparities within socio-economic groups. The analysis will proceed in three steps. First, the observed statistical relationship between development levels and income inequality is briefly discussed for a sample of countries and periods where comparable data are available. Second, a more structural approach is discussed where additional variables representing in some way the socio-economic group structure of the population are introduced. Finally, semi-structural studies of the evolution of the distribution in selected countries are discussed. Interestingly enough, the conclusions obtained about the effect of growth on inequality varies rather radically from one approach to another.

Correlation between Growth and Inequality: Rise and Fall of the Kuznets Curve

As illustrated in the bottom table 1 above, the cross-country evidence on the distributional effects of growth is essentially inconclusive. It is true that a cross-section of countries taken in the 1970s suggests a parabolic relationship that seems in agreement with Kuznets' hypothesis. But cross-sections taken at a later date, and presumably with better data, fail to yield statistically significant results. More importantly, when controlling for country fixed effects so as to isolate the average consequences of within country growth on distribution, no statistically significant effect can be identified either.

The preceding results fit well the existing literature on the distributional consequences of growth and the Kuznets curve. Back in the 1970s and in the early 1980s, several papers provided pure cross-sectional estimates of the distributional consequences of growth that seemed in agreement with Kuznets' hypothesis, with very much emphasis on the estimation of the turning point at which further growth would cause inequality to go down rather than up – see in particular Paukert (1971), Ahluwalia (1974, 1976a and b), Lecaillon et al. (1979). This early literature was very much criticized for its lack of econometric rigor and the quality of the data being used - see in particular Anand and Kanbur (1991). When better and more complete data became available, it indeed turned out that the parabolic shape of the relationship between income and inequality was a feature of the 1970s. This feature vanished with the data available in subsequent periods, whereas the first attempts at controlling for country fixed effects confirmed that the findings based on the observations of the 1970s were not robust – see Anand and Kanbur (1993), Fields and Jakubson (1993) and Deiniger and Squire (1998), probably the most data-comprehensive analysis available although it relies on secondary data sources.

The interest in the Kuznets hypothesis has not completely vanished today and this hypothesis is frequently revisited in the light of new data and estimation techniques. But the cottage industry that developed around trying to confirm or reject this hypothesis is in

decline.[33] At this stage, it seems fair to say that a consensus has emerged according to which available data *do not suggest any strong and systematic relationship* between inequality and the level of development of an economy. Even those authors who identified a significant relationship agree that it is weak and explains little of the observed differences in inequality.[34]

As in the preceding sections, it may be worth exploring how time series compare to cross-country differences in terms of the correlation between income inequality and development levels. Indeed, considerable attention has been given lately to the evolution of inequality and to the question of whether there was a systematic tendency for modern growth to generate more inequality, as observed in a few countries during the 1980s. This literature is mostly concerned with developed countries because of data availability, even though time series of distribution data in those countries are not always consistent - see Atkinson and Brandolini (2001). The common conclusion of most existing analyses is that inequality has substantially increased in a number of countries between the mid 1980s and the mid 1990s - 12 out of the 17 countries analyzed by Gottschalk and Smeeding (2000). Yet, this evolution must be contrasted with the fact that inequality had been declining in almost all developed countries throughout the 1960s and the 1970s, so that inequality in many countries today is comparable to what it was 30 years ago. It is also worth stressing that inequality failed to increase in a few countries, most likely thanks to very efficient -and possibly increasingly so - redistribution. Time series in developed countries thus seem to confirm the evidence based on cross-country analysis,

[33] No reference has been made here to the historical literature on the Kuznets curve. It generally points to sizable changes in some particular inequality measures over long periods of time, which sometimes conform with Kuznets' own finding. Yet, the problem is that it relies on very rough measures of inequality – see below p. 46.

[34] See for instance the conclusion reached by Barro (2000) and his reference to Papanek and Kyn (1986) for a similar statement. No mention has been made here of the few studies focusing on the effect of the rate of growth, rather than the level of economic development on inequality. Lundberg and Squire (2003) and Banerjee and Duflo (2003) find that faster economic growth tends to be associated with increasing inequality, but here again the effect is extremely small. Note finally that, after raising the same hopes as the Kuznets hypothesis in the 1970s and 1980s, the recent literature on the reverse causation, seems also to converge to the conclusion of no 'systematic' effect of inequality on growth. This was pointed out by Benabou (1996) and has been confirmed since then by contradictory results obtained with different specifications and samples - see Banerjee and Duflo (2003) .

namely that there is no significant long-run trend, related to the growth process or not, affecting the degree of within-country inequality.

Although data are still more shaky, the same conclusion seems to hold for middle- and low-income countries. Cornia (2001) and Cornia, Addison and Kiiski (2004) find that, out of 34 developing countries for which they have several observations between the 1950s and the mid 1990s, inequality is higher in the terminal period for 15 of them, equal for 14 and lower for 5. Yet, little is said about the intermediate years. When data are available, a U-shape evolution is observed in a number of cases where inequality is found to be increasing when comparing the terminal and the initial years. Overall, clear ascending or descending trends over long periods of time thus are infrequently observed.

Dealing with the effect of growth on social structures, one might prefer to use the concept of poverty, which may have a firmer social connotation than inequality. With this concept, a dimension of social structures is simply the poor-non-poor difference, that is the proportion of people living in poverty and the distance at which those people find themselves from the poverty line and from the average income of the non-poor. Of course, if income is taken as the only dimension of welfare, poverty and inequality are rather equivalent concepts when poverty is defined in relative terms, as for instance the proportion of people living with less than 50% of the median income. Poverty, then, becomes a particular inequality measure and much of the preceding argument presumably applies to poverty as well as inequality. Things are different when poverty is defined in absolute terms, as with the widely used international poverty line of $1 a day or any national poverty line representing the minimum budget deemed necessary for survival. Growth then plays a direct role to explain the evolution of poverty. In particular, if the distribution of relative incomes remains constant over time, then changes in poverty essentially reflect uniform income growth in the population. On the contrary, when distribution changes over time, economic growth may play a more complex role depending on its actual effects on distribution.[35] That cross-country analysis fails to find

[35] On the identity that relates income distribution, poverty and growth see Fields (2001), Ravallion and Datt (2002) or Bourguignon (2003).

any systematic relationship between distribution and growth suggests indeed that growth has the simpler and more direct impact on poverty that was just mentioned. This is essentially the argument that led Dollar and Kraay (2002) to conclude that 'growth is good for the poor'. However, it will be seen below that this conclusion may hide considerable disparities across countries.

Towards structural estimates of the effects of growth on distribution ?

By focusing on the correlation between inequality and development levels, both in cross-section and time series, the preceding section does not do justice to existing work. Most empirical models that may be found in the literature actually comprise additional variables that may explain the evolution of inequality alongside with, or independently from growth. For instance, the regressions in the pioneer paper by Alhuwalia (1976b) had the income share of various quantiles on the left-hand side and a host of variables on the right-hand side, together with GDP per capita and its square. These variables included the GDP share of agriculture, some educational indicators and some fertility indicators. Thus, the effect of the development level of inequality was supposed to come on top of the effects of sectoral shifts, and associated changes in relative factor rewards. In terms of the analysis in this chapter, the implicit objective of such a specification is somewhat unclear. The distribution between socio-economic groups implicitly defined by variables like the GDP share of agriculture or the degree of urbanization and its impact on overall inequality is taken care of precisely by the presence of these variables in the regression. Under these conditions, the purpose of keeping GDP per capita among the regressors could only have been to identify the effect of economic growth on the distribution of income *within* those groups, a rather restrictive objective. Retrospectively, the economic framework behind those regressions thus appears as essentially ad hoc, between the reduced form model implicit in the regressions in table 1 and a true structural model identifying the channels through which growth may indeed affect the distribution. Unfortunately, this imprecision of the economic framework behind the regressions being estimated is rather frequent in the empirical literature on inequality and development.

Using structural forms in cross-country regressions is possible, but probably requires gathering the appropriate data. In Bourguignon and Morrisson (1990, 1998), for instance, the objective is to explain cross-country differences in distribution explicitly through a model resembling (2) above but with a larger number of assets. Agent ”'s income is thus taken to be given by :

$$y_i = \sum_j a_{ij} w_j F_j$$

where a_{ij} is the share of factor j owned by agent i, the total endowment of factor j is given by F_j, and w_j is its remuneration rate. Endogenizing the latter within the framework of a small open economy, it is then possible to write the overall distribution of income $\{y\}$ as:

$$\{y\} = H\left[\{a_{.j}\}, F_{.}, p, t\right] \qquad (8)$$

where the arguments of the function $H(\,)$ are the distribution of resources in the population, $\{a_{.j}\}$, the vector of total endowments, $F_{.}$, the vector of international prices faced by the economy, p, and the tax rates and tariffs that it imposes, t. Then, one may try to proxy for these various arguments by aggregate data available on a comparable basis across countries and use the resulting empirical model to analyze the effects of economic growth on the distribution.

The two papers referenced above stop short of the latter objective, mostly because of the difficulty of identifying all the variables necessary for the analysis and the necessity of using very imperfect proxies, which prevent proceeding with a truly structural analysis. For instance, physical capital per worker is approximated by GDP per capita. This is not unjustified for a given value of other aggregate endowments of productive factors, but this makes it impossible to distinguish the distributional effects of capital accumulation and of total factor productivity. Even so, however, the analysis shed light on the effects of the ownership distribution variables - land, human capital - on the distribution of income at one point of time, as well as on the impact of relative aggregate endowments (land,

physical and human capital and raw labor) and policy variables like trade protection. In line with the argument in the previous sections, it also showed the importance of labor market competitive imperfection, and in particular the 'dualism' of the economy as represented by the relative productivities of the agricultural and non-agricultural sectors.

Going beyond these partial results and analyzing the potential effects of growth in the distribution of income within the framework of (8) is still on the agenda. Better data are necessary. As rough as existing empirical applications may be, however, this analysis suggests that the distributional effects of growth are complex and most likely to be strongly differentiated across countries. With the preceding specification and in view of the results obtained, it appears clearly that growth affects the distribution through various channels - changes in relative factor endowments, changes in the distribution of these factors, changes in policies, changes in the functioning of factor markets – and in a way that may depend on the initial value of these various macro-economic characteristics.

Case Study Analysis

Should one then conclude with Dollar and Kraay (2002) that growth is distributionally neutral and therefore that 'growth is good for the poor', whatever the engine behind it? The preceding argument suggests that it would be going too far. What the important literature on the effect of growth on inequality shows is essentially that there is apparently no significant relationship that would be valid across countries and time periods. It certainly does not say that this is true for specific countries during a particular period. In effect, sizable changes have been observed in several countries in the recent past, the causes of which are not always readily apparent but which are not necessarily independent from economic growth and some of its features.[36] In effect, most case studies on the evolution of inequality over time single out characteristics of the growth process or of policies behind it that are responsible for specific distributional changes.

[36] See for instance Atkinson and Brandolini (2002) for developed countries and Ravallion and Chen (1997) for developing countries.

The debate that took place recently on whether recent increases in earnings inequality in various developed countries were due to technical progress and the consequent shift of developed economies towards high-tech is a good example of such an approach.

An important stream of case studies on the distributional consequences of growth is found in the historical literature. Following the example of Kuznets himself, numerous economic historians tried to identify the trend in some inequality measure over long periods of time. Findings often conform with the Kuznets curve hypothesis. Yet, a common problem is that very much of that literature relies on very rough measures of inequality often based on a few macro-economic characteristics and ignores important sources of micro-economic heterogeneity. Because of this, it tends to over-emphasize the role of phenomena like urbanization or the shift away from self-employment, that is sectoral shift phenomena. A survey of findings is offered by Lindert (2000) and Morrisson (2000).

Identifying empirically the forces through which economic growth may shape the distribution of income in actual growth experiences is a difficult task because it requires correcting the observed evolution of the distribution for sources of change unrelated or very loosely related to growth. An exercise of this type was undertaken in a series of case studies that explore the microeconomics of income distribution dynamics (MIDD) in a small number of middle-income countries.[37] The following example illustrates the difficulty of empirically isolating the distributional effects of growth and shows at the same time the major potential role that growth plays in distributional issues.

The methodology used in the MIDD study consists in decomposing the observed change in the distribution of earnings or per capita income into three types of effects, which parallel the general argument in this chapter. The first type corresponds to changes in the structure of earnings for given socio-demographic characteristics of individuals, and possibly by labor market segment if the labor market is imperfectly competitive. The second set of effects corresponds to a change in the labor supply of individuals or

[37] Bourguignon, Ferreira and Lustig (2004)

household members, and possibly their allocation across labor market segments. The final set of effects includes those occurring due to a change in the distribution of socio-demographic characteristics of individuals and households. In terms of the analysis in this chapter, the second and third type of effects would, in some sense, correspond to the shift of people across socio-demographic groups, whereas the first one would correspond to changes in income differentials across these groups, a residual term actually included in the third effect, accounting for changes in the distribution within groups. Each effect is estimated by simply importing in the initial year the features of the final year in one of the various dimensions just indicated, or vice-versa. Thus, the effect of the structure of earnings is obtained by simulating what would be the distribution of income in year *1* if the structure of earnings by socio-demographic characteristics (gender, education, region, etc) had been the one observed in year *2*. The 'fertility effect' is obtained by importing in the initial year the same relationship between family size and parents' characteristics (education, age, race, region, etc) as the one observed in the terminal year, etc… [38]

This methodology has been applied, among other countries, to Brazil during the period 1976-1996.[39] What is remarkable during that period in Brazil is that neither the mean income – or GDP per capita – of the Brazilian population nor inequality changed much, even though most usual aggregate inequality measures show a moderate decline. From this direct observation, one might then conclude that very slow growth was associated in Brazil with virtually no change in the distribution of earnings or income. This would be erroneous, however. What happened is that other phenomena compensated for the distributional effects of slow growth. The decomposition methodology presented above led to 3 points. a) Over the period under analysis, family size went down significantly and more so among people with low education and income. Because of this factor, inequality in Brazil should have substantially declined. b) The structure of earnings changed moderately against least-skilled and self-employed workers. c) It also turned out that the occupational structure of the population was modified. Employment in general, and employment in the formal sector in particular, had gone down, hitting more severely

[38] This methodology is explained in detail in chapter 2 of Bourguignon Ferreira and Lustig (2004). See also Bourguignon, Fournier and Gurgand (2001).
[39] See Ferreira and Paes de Barros (1999, 2004).

the segments of the population with the lowest education levels and at ages where access to the labor market is the least easy – young and old people. Overall, however, these various changes tended to compensate each other.

Although the methodology described above does not include any formal representation of the labor market and the way it may be affected by growth, it is difficult not to relate points b) and c) to the sluggish growth performance of Brazil during the two decades under analysis. Within a dual economy framework, which seems to fit well the Brazilian economy, the general story would thus be as follows. Slow growth was responsible for a weak labor market, which may have caused an increasing skill differential in the earnings of wage workers and self-employed, as well as job losses or worker discouragement among the least skilled. Both phenomena, but mostly the latter, actually contributed to an increase in inequality. The reason why this inequality increase did not actually show up is that it was compensated by falling family sizes which were more pronounced at the bottom of the distribution - and to a lesser extent progress in the education level of the poorest. Slow growth in Brazil might thus have been responsible for increased inequality after all.

This example shows that identifying the actual effects of economic growth on the social structure of a population may require more than simply observing the changes in that structure and the rate of growth during a given time period. Some of the observed change in social structures may indeed be directly imputed to what is happening on the economic growth front, but may also be due to other concomitant phenomena which are independent of growth or very indirectly related to it. With this example in mind, one may then understand perfectly that the phenomena put forward by economic theory to explain how economic growth is likely to affect social structures may be difficult to observe in actual growth episodes. This is because parallel phenomena, not directly related to growth - but not necessarily independent from it either - affect the distribution and introduce some noise in the observation, this being true both in cross-sectional and time series analyses.

The important point here is that the channels identified by theory may well prove empirically relevant when the necessary correction of available data for other non-neutral distributional phenomena has been made. Of course, the difficulty is to know whether these phenomena are themselves growth dependent or not. In the preceding case, did the change in fertility take place in an autonomous way, or was it the result of economic growth per se, or possibly the result of an educational improvement which itself could have been autonomous or the result from growth? It is this kind of structural model that must be confronted with the data, rather than the very reduced form models behind cross-sectional work or simple comparisons of changes in inequality and income per capita measures. This is a much more difficult exercise, which has barely begun.[40]

4. CONCLUSIONS

Some strong conclusions emerge from our review of the literature on the consequences of economic growth for social structures. They may be summarized as follows.

(a) The Foremost Importance of Sectoral Shift Phenomena

If we take a historical or a cross-country perspective, it is clear that the major consequences of economic growth for social structures go through sectoral shifts of the population, and possibly, at a later stage, through shifts across more narrowly defined socio-economic groups. The shift from traditional agriculture to modern agriculture or non-agricultural activities in the urban sector is at the heart of the social and cultural history of most industrialized countries. It is still today at the heart of the social evolution in countries that are 'emerging' as well as in those at a much earlier stage of development. The sectoral shift is also present in so-called advanced countries that have started their journey towards the 'post-industrialization' stage where most activity will increasingly be directly or indirectly linked to services of various types rather than manufacturing.

[40] Steps in that direction have been taken in the recent 'micro-macro' literature. See in particular Browning, Hansen and Heckman (1999), Bourguignon and Pereira da Silva (2003, introduction), Heckman, Lochner and Taber (1998), Townsend and Uema (2001)

Transforming factory workers into administrative employees involves major changes in social structure, just like transforming small farmers into industrial workers.

Other important shifts take place throughout the growth process, the most powerful ones from a social point of view being the drive towards education and a skilled labor force and a fuller integration of women into market activities.

(b) The Role of Market Integration

It was observed historically that economic development proceeds together with a 'marketisation' of societies. Markets develop by covering an increasing share of all transactions. The most significant difference between sectoral shifts observed in industrialized countries on the one hand and in low- or lower-middle income countries on the other is not only the nature of the sectors involved but also the degree of market integration of society. Sectoral shifts at early stages of economic growth take place within markets that are functioning very imperfectly. Because of this, they have a strong impact on society, radically changing relative income levels and income-related behaviors. At later stages of development, markets function better and sectoral shifts take place in a smoother manner. At the limit, some sectors may lose weight in favor of others with only small changes taking place in the structure of earnings and in the social features associated with it. With such a market integration, the real issue is the speed of adjustment to growth, i.e., whether the structure of the demand for productive factors changes faster than the supply of these factors, modifying in one direction or another the structure of remunerations of these factors. Thus, economic growth in a market integrated economy has social consequences different from less integrated economies. In the former case, economic growth may take place in a somewhat 'balanced' way without any marked impact upon social structures and, in particular, on the distribution of well-being within the population. This is what is implicitly assumed in many contemporaneous models.

(c) Social Costs of Transitory Adjustment

In reality, such balanced growth paths are unlikely to be observed. In presence of any permanent shock in technology or in policy, adjustment lags in either the demand or the supply side of factor markets produce tensions with possibly durable social effects. Case studies of both developing and developed countries reveal the presence of such tensions. Unemployment problems in some European countries or the rise in earnings inequality in the US during the 1980s and 1990s are sometimes seen as symptoms of such tensions. Even though these disequilibria may be expected to be transitory, they may have powerful and long-lived social consequences, potentially able to affect future economic growth. A fortiori, this conclusion holds true when markets mechanisms are working imperfectly.

Some other dimensions of social structures have been examined in this chapter that fit these conclusions or may be direct consequences of them. This is the case in particular of the economic role of women and male-female economic and social differences. Sectoral shifts explain the changing roles of men and women, whereas the move towards market integrated economies may be the cause of increased female labor force participation, as well as of decreasing male-female differences both in the economic and the social spheres.

These are important conclusions. Yet, as indicated at the beginning of this chapter, this is only part of the whole story. Two other parts are missing. The first is the second half the circle that links economic growth and social structures. By choice, this chapter has focused on the social consequences of growth, but the effect of changes in social structures on the pace and structure of economic growth is equally important—as shown in various chapters of this Handbook. The second missing part has to do with the effect of changes in social structures on social institutions themselves, and on social relations. This is briefly discussed in closing this chapter.

(d) Effect of Growth on Social Structures through Social Institutions

In addition to the changes in social structures resulting directly from economic growth, the general way in which individuals relate with each other is also likely to be modified under the effect of growth. For instance, a more affluent population may have more time and interest for collective tasks and for political participation, thus modifying the nature of the public decision-making process. As another example, a higher standard of living may induce a higher risk aversion and therefore a demand for more social insurance. We conclude this chapter by focusing on two examples of how social relations are likely to be modified under the effect of economic growth, with the social structure effect analyzed above often playing an intermediate role in that evolution.

The first example has to do with the size of the public sector. It is the well-known "law of expanding state activities" formulated in 1883, in the midst of a period of rapid industrialization and social change, by Adolf Wagner (Wagner 1883, Cameron 1978). Several explanations have been proposed to explain this law, some of them directly linked to growth and some others indirectly, through the channels of increasing education and political participation.[41] Theories of the direct link between economic growth and government size emphasize the income elasticity of the demand for certain types of public goods, such as roads or education, or for the correction of negative public externalities due to growth, such as pollution or congestion. Theories of the indirect link insist on the redistribution role of government and the way in which growth may affect the factors that determine that role. For instance, in the classic Meltzer and Richard (1981) paper, the growth of government is explained by the combination of the expansion of universal suffrage and initial inequality in the distribution of income. Then the drive towards democracy is explained by the effect of economic growth on political participation either through a direct income effect [42], or through an indirect one as in the political economy models proposed for instance by Justman and Gradstein (1999) or

[41] These different theories regarding the size of government are reviewed by Lybeck and Henrekson (1988), who focus on empirical tests of these theories, and Mueller (2003).

[42] See for instance Frey (1971, 1972), Huntington and Nelson (1976), Gradstein and Justman (1995).

Acemoglu and Robinson (2000). A second channel is education brought about by, or coming together with growth.[43]

It is difficult to test for the relationship between economic growth and government size while taking into account the various channels mentioned above. In a cross-country reduced form framework, however, it is interesting that the relationship is a rather weak one, especially when one controls for fixed county effects – see table 1.[44]

As a second, and not unrelated example, consider the demand for social insurance. As for Wagner's law, there are direct and indirect effects of growth on social insurance and family structures. Direct effects may go through risk aversion or idiosyncratic risks themselves increasing with the income level, or through richer societies being able to afford the fixed costs of social insurance "technology" (related to monitoring of income, contribution collection, management and information systems and other organisational costs of social insurance).[45] An indirect effect of growth on social insurance is via urbanization and the consequent phasing out of high fertility and extended household arrangements throughout societies. This creates a demand for a substitute to the insurance provided by the extended family system in poorer societies.

These two examples about the way social institutions may be affected directly or indirectly by economic growth show the power of economic growth and its determinants to transform the social functioning of societies and not only their economic characteristics. Of course, these changes in the way individuals relate to each other, the way they make public decisions and the nature of these decisions come on top of all other effects of growth. These include changes in consumption behavior caused by increasing income and technical progress as well as the changes in social structures analyzed throughout this chapter. More than in the latter cases, however, changes in social

[43] See the sociological literature on political participation, for instance Brady, Verba and Shlozman (1995). See also the formal analysis of the link between democracy, education and growth in Bourguignon and Verdier (1999).

[44] Some authors also postulate an inverted U-curve relationship between government size and income level. See Grossman (1987, 1988) and Peden (1991).

[45] As in the modeling of financial development in Saint-Paul (1992) or Tressel (2003).

institutions are affected by a host of non-economic phenomena that make it difficult to identify and isolate the role of economic growth. Economic determinism might be dangerous here. In-depth case studies combining the whole range of social sciences would seem the appropriate way of approaching this issue.

REFERENCES

Acemoğlu Daron (2002). "Technical Change, Inequality and the Labor Market," *Journal of Economic Literature*, 40(1), pp. 7-72

Acemoğlu, Daron and James A. Robinson (2000). "Why Did the West Extend the Franchise? Democracy, Inequality and Growth in Historical Perspective.", *Quarterly Journal of Economics*, 115(4), pp. 1167-99

Acemoğlu Daron and James Robinson (2000). "A Theory of Political Transitions", MIT Working Paper.

Adelman, Irma, and Cynthia Taft Morris (1967) *Society, Politics and Economic Development- A Quantitative Approach*. Baltimore: Johns Hopkins University Press.

Adelman, Irma and Sherman Robinson (1988), Income Distribution and Development, In H. Chenery and T.N Srinivasan (eds) *Handbook of Development Economics*, Vol I, NorthHolland, Amsterdam

Aghion, Philippe (2002), Schumpeterian Growth Theory and the Dynamics of Income Inequality, *Econometrica*, 70(3), pp. 855-82

Aghion, Philippe, and Patrick Bolton (1997). "A Trickle-down Theory of Growth and Development." *The Review of Economic Studies* 64:2, pp. 151–172, April.

Aghion, Philippe, Eve Caroli, and Cecilia Garcia-Peñalosa (1999). "Inequality and Economic Growth: The Perspective of the New Growth Theories." *Journal of Economic Literature* 37:4, pp. 1615-1660, December

Aghion, Philippe and Peter Howitt (1998), *Endogenous Economic Growth Theory*, The MIT Press

Aghion, Philippe, Peter Howitt and Giovanni Violante (2002), General Purpose Technology and Wage Inequality, *Journal of Economic Growth*, 7(4), pp. 315-45

Ahluwalia, Montek (1976a). "Income Distribution and Development: Some Stylized Facts" *American Economic Review* 66:2, pp. 128-135, May.

Ahluwalia, Montek (1976b), Inequality, Poverty and Development, Journal of Development Economics, 3: pp. 307-342.

Ahluwalia, Montek (1974). "Income Inequality: Some Dimensions of the Problem." *Finance and Development* 11:3, pp. 2-8, September.

Alberto Alesina and Edward Glaeser (2004). *Fighting Poverty in the US and Europe*. Oxford: Oxford University Press

Anand, Sudhir, and Ravi Kanbur (1991), Inequality and Development : a Critique, Journal of Development Economics, 41, pp. 19-43

Anand, Sudhir, and Ravi Kanbur (1993). "The Kuznets Process and the Inequality-Development Relationship." *Journal of Development Economics* 40:1, pp. 25-52, February.

Arellano, Manuel, and Stephen Bond (1991). "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *The Review of Economic Studies* 58:2, pp. 277-297, April.

Asbell, Bernard (1995). *The Pill: A Biography of the Drug That Changed the World*. New York: Random House.

Ashenfelter, Orley C. and David Card, eds. (1999), *Handbook of Labor Economics*, 3 volumes. Amsterdam: North-Holland

Atkinson, Anthony B. (1999), *The Economic Consequences of Rolling Back the Welfare State*, Cambridge: The MIT Press.

Atkinson, Anthony B., and Andrea Brandolini (2001). "Promise and Pitfalls in the Use of 'Secondary' Data-Sets: Income Inequality In OECD Countries as a Case Study." *Journal of Economic Literature* 39:3, pp. 771-99, September.

Atkinson, Anthony, Lee Rainwater and Timothy Smeeding (1995), Income Distribution in OECD Countries : Evidece form the Luxembourg Income Study, OECD, Paris

Bacha, Edmar L., and Lance Taylor (1978). "Brazilian Income Distribution in the 1960s: 'Facts,' Model Results and the Controversy." *Journal of Development Studies* 14, pp. 271-297, April.

Baldwin, Robert E. and Glen G. Cain (2000), "Shifts in relative U.S. wages: the role of trade, technology, and factor endowments," *Review of Economics and Statistics*, 82(4):580-95, November

Banerjee, Abhijit, and Esther Duflo (2003). "Inequality and Growth: What Can the Data Say?" *Journal of Economic Growth* 8:3, pp. 267-299, September.

Banerjee, Abhijit V., and Andrew Newman (1993). "Occupational Choice and the Process of Development." *Journal of Political Economy* 101:2, pp. 274–298, April.

Barro, Robert J. (1997), "Determinants of Democracy," Harvard Institute of International Development Discussion Paper No.570, January 1997.

Barro, Robert J. (2000). "Inequality and Growth in a Panel of Countries." *Journal of Economic Growth* 5:1, pp. 5-32, March.

Baumol, William J., Alan S. Blinder, and Edward N. Wolff (2003). *Downsizing in America: Reality, Causes, and Consequences*. Russell Sage Foundation.

Benabou, Roland (1996). "Inequality and growth." *NBER Macroeconomics Annual 1996*, pp. 11-76.

Bertola, Guiseppe (1993). "Factor Shares and Savings in Endogenous Growth." A*merican Economic Review* 83:5, pp. 1184-1198, December.

Birdsall, Nancy, and Lauren A. Chester (1987). "Contraception and the Status of Women: What Is the Link?" *Family Planning Perspectives* 19:1, pp. 14-18, January–February.

Blau, Francine and Lawrence Kahn (1997), Swimming Upstream: Trends in the Gender Wage Differential in the 1980s, Journal of Labor Economics, 15:1, pp. 1- 42

Blau, Francine and Lawrence Kahn (2003), Understanding International Differences in the Gender Pay Gap, In Brigida Garcia, Richard Anker, and Antonella Pinnelli, eds., *Women in the Labour Market in Changing Economies: Demographic Issues*. Oxford: Oxford University Press, pp..

Brady Henry, Sidney Verba and Kay Schlozman (1995), Beyond SES: a Resource Model of Political Participation, *American Political Science Review*, 89(2), pp. 271-94.

Blundell, Richard, and Stephen Bond (1998). "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models." *Journal of Econometrics* 87:1, pp. 115-143,

Bourguignon, François (1981). "Pareto Superiority of Unegalitarian Equilibria in Stiglitz' Model of Wealth Distribution with Convex Saving Function." *Econometrica* 49:6, pp. 1469-1475

Bourguignon, François (1990). "Growth and Inequality in the Dual Model of Development : The Role of Demand Factors." *The Review of Economic Studies* 57:2, pp. 215-228, April.

Bourguignon, François (2003), "The Growth Poverty Inequality Triangle," Processed, World Bank, Washington DC.

Bourguignon, François and Christian Morrisson (1990), Income Distribution, Development and Foreign Trade: A Cross-Sectional Analysis, *European Economic Review*, 34, 1990, pp. 1113-1132.

Bourguignon, François and Christian Morrisson (1998), Inequality and Development: the Role of Dualism, *Journal of Development Economics*

Bourguignon, François, Martin Fournier and Marc Gurgand (2001), Fast development with a stable income distribution : Taiwan, 1975-1994, *Review of Income and Wealth*, 47(2), 139-63

Bourguignon, François and Luiz Pereira da Silva, editors (2003). *The Impact of Economic Policies on Poverty and Income Distribution*, A Co-publication of the Wortld bank and the Oxford University Press.

Bourguignon, François, Francisco Ferreira, and Nora Lustig, eds. (2004). M*icroeconomics of Income Distribution Dynamics in East Asia and Latin America*., Oxford University Press and The World Bank, Washington, DC

Browning, Martin, Lars Peter Hansen, James J. Heckman (1999). "Micro Data and General Equilibrium Models." In Taylor, J. and M. Woodford (eds.), *Handbook of Macroeconomics*, vol. 1A. Amsterdam: North-Holland.

Card, David (1999). "Causal Effect of Education on Earnings," in: Ashenfelter, O. and D. Card, eds., *Handbook of Labor Economics*, vol. 3A, pp. 1801-1863, Amsterdam: North-Holland

Cameron, David (1978). "The Expansion of the Public Economy: A Comparative Analysis," *American Political Science Review*, vol. 72(4), pp. 1243-1261, December.

Clark, Robert, Anne York, and Richard Anker (2003). "Cross-national Analysis of Women's Labour Force Activity Since 1970." In Brigida Garcia, Richard Anker, and Antonella Pinnelli, eds., *Women in the Labour Market in Changing Economies: Demographic Issues*. Oxford: Oxford University Press, 2003, pp. 13-34.

Clemens, Michael (2004), "The Long Walk to School: International Education Goals in Historical Perspective," Center for Global Development Working Paper No. 37, March

Chenery, Hollis, Montek S. Ahluwalia, C.L.G. Bell, John H. Duloy, and Richard Jolly (1974). *Redistribution with Growth: Policies to Improve Income Distribution in Developing Countries in the Context of Economic Growth : A Joint Study*.

Commissioned by the World Bank's Development Research Center and the Institute of Development Studies, University of Sussex. Oxford: Oxford University Press.

Chenery, Hollis B., and Moises Syrquin (1975). *Patterns of Development, 1950-1970*. London : Oxford University Press for the World Bank.

Cornia, Giovanni Andrea with S. Kiiski (2001), Trends in Income Distribution in the Post World War II Period : Evidence and Interpretation, UNU/WIDER Discussion Paper 89, Helsinki, Finland

Cornia, Giovanni Andrea, Tony Addison and Sampsa Kiiski (2004), "Income Distribution Changes and their Impact in the Post-Second World War Period," in: G. Cornia ed., *Inequality, Growth and Poverty in an Era of Liberalization and Globalization*, Oxford: Oxford University Press.

de Janvry, Alain, and Elizabeth Sadoulet. (1983). "Social Articulation as a Condition for Equitable Growth." *Journal of Development Economics* 13:3, pp. 275–303, December.

Deininger, Klaus, and Lyn Squire (1998). "New Ways of Looking at Old Issues: Inequality and Growth." *Journal of Development Economics* 57:2, pp. 259-287.

Dollar, David, and Art Kraay (2002). "Growth is Good for the Poor." *Journal of Economic Growth* 7:3, pp. 195-225, September.

Durand, John (1975). *The Labor Force in Economic Development: A Comparison of International Census Data, 1946-66*. Princeton, NJ: Princeton University Press.

Durlauf, Steve and William Brock (2001), Growth Empirics and Reality, World Bank Economic Review, 15, 229-272

Easterlin, Richard A. (1996). *Growth Triumphant: The Twentieth Century in Historical Perspective*. Ann Arbor, MI: University of Michigan Press.

Eicher, Theo (1996), Interactions between Endogenous Human Capital and Technical Change, *Review of Economic Studies*, 63(1), pp. 127-44

Fei, John C. H., and Gustav Ranis (1965). "Innovational Intensity and Factor Bias in the Theory of Growth." *International Economic Review* 6:2, pp. 182-198, May.

Fei, John C. H., and Gustav Ranis (1964). *Development of Labor Surplus Economy Theory and Policy*. Homewood Illinois: Richard D Irwin.

Ferreira, Francisco and Ricardo Paes de Barros (1999), The Slippery Slope: Explaining the Increase in Extreme Poverty in Urban Brazil, 1976-1996, *Revista de Econometria*, 19 (2), pp.211-296

Ferreira, Francisco and Ricardo Paes de Barros (2004), Climbing a Moving Mountain: Explaining the Decline in Income Inequality in Brazil from 1976 to 1996, in F. Bourguignon, F. Ferreira and N.Lustig (eds), M*icroeconomics of Income Distribution in East Asia and Latin America*, Oxford University Press and the World Bank.

Fields, Gary S. (1979). "A Welfare Economic Approach to Growth and Distribution in the Dual Economy." *The Quarterly Journal of Economics* 93:3, pp. 325-353, August.

Fields, Gary S., and George H. Jakubson (1994). "New Evidence on the Kuznets Curve." Cornell University, mimeo.

Fields, Gary S. (2001), *Distribution and Development: A New Look at the Developing World*, Russell Sage Foundation and the MIT Press

Frankel, Marvin (1962). "The Production Function in Allocation and Growth: A Synthesis," *American Economic Review*, 52:5, pp. 995-1022.

Frey, Bruno (1971), Why Do High Income People Participate More in Politics?, Public Choice, 11, pp. 100-05

Frey, Bruno (1972), Political Participation and Income Level, Public Choice, 13, pp.119-22

Galor, Oded, and Joseph Zeira (1993). "Income Distribution and Macroeconomics." *The Review of Economic Studies* 60:1, pp.35-52, January.

Galor, Oded, and Daniel Tsiddon (1997), Technological Progress, Mobility and Economic Growth, *American Economic Review*, 87(3), pp. 363-82

Goldin, Claudia (1995). "The U-Shaped Female Labor Force Function in Economic Development and Economic History." In T. P. Schultz, ed., *Investment in Women's Human Capital and Economic Development*. Chicago: University of Chicago Press.

Goldin, Claudia (1991). "The Role of World War II in the Rise of Women's Employment." *American Economic Review* 81:4, pp. 741-56, September.

Goldin, Claudia, and Lawrence F. Katz (2002). "The Power of the Pill: Oral Contraceptives and Women's Career and Marriage Decisions." *Journal of Political Economy* 110:4, pp. 730-770, August.

Goodin, Robert, Bruce Headey, Ruud Muffels and Henk-Jan Dirven (1999). *The Real Worlds of Welfare Capitalism*, Cambridge: Cambridge university Press.

Gottschalk, Peter and Timothy Smeeding (2000), Empirical Evidence on Income Inequality in Industrialized Countries, In A. Atkinson and F. Bourguignon (eds), Handbook of Income Distribution, Vol. I, North-Holland, Amsterdam, pp. 261-297

Gradstein, Mark and Moshe Justman (2002). "Education, Social Cohesion, and Economic Growth," *American Economic Review*, vol. 92(4): 1192-1204.

Grossman, Philip J. (1987), "The Optimal Size of Government," *Public Choice* 53, pp. 131-47.

Grossman, Philip J. (1988), "Government and Economic Growth: A Nonlinear Relationship," *Public Choice* 56, pp. 193-200.

Gunderson, Morley (1989), Male-Female Wage Differentials and Policy Responses, Journal of Economic Literature, 27:1, pp. 46-72

Heckman, James J. (2001). "Micro Data, Heterogeneity, and the Evaluation Of Public Policy: Nobel Lecture." *Journal of Political Economy* 109: 4, pp. 673-748, August.

Heckman, James J., Lance Lochner and Eric Traber (1998), "Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents," *Review of Economic Dynamics*, vol. 1, p. 1-58.

Juhn, Chinhui, Kevin Murphy and Brooks Pierce (1993), Wage Inequality and the Rise in Returns to Skill, Journal of Political Economy, Vol. 101(3); p. 410-443.

Justman, Moshe and Mark Gradstein (1999). "The Democratization of Political Elites and the Decline in Inequality in Modern Economic Growth." in Brezis E.S. and P. Temin, editors, *Elites, Minorities and Economic Growth*. New York: Elsevier, North-Holland.

Kanbur, Ravi (2000) Income Distribution and Development, In A. Atkinson and F. Bourguignon (eds), *The Handbook of Income Distribution*, Vol 1, North-Holland, Amsterdam

Katz, Lawrence F., and David H. Autor (1999). "Changes in the Wage Structure and Earnings Inequality." In O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics*, vol. 3A, pp. 1463-1555. Amsterdam: North-Holland.

Katz, Lawrence and Kevin Murphy (1992), Changes in Relative Wages, 1963-1987: Supply and Demand Factors, *Quarterly Journal of Economics*, 107(1), pp. 35-78

Knight, J. B. (1976), Explaining Income Distribution in Developing Countries : a Framework and an Agenda, *Oxford Bulletin of Economics and Statistics*, 38(3), pp.161-77

Krueger, Alan and Mikael Lindahl (2001), "Education for Growth: why and for whom?" Journal of Economic Literature, vol. 39(4): 1101-1136

Kuznets, Simon (1955). "Economic Growth and Income Inequality." *American Economic Review* 45:1, pp. 1-28, March.

Kuznets, Simon (1966), *Modern Economic Growth : Rate, Structure and Spread*, Yale University Press, New Haven

Lee, Ronald (2003). "The Demographic Transition: Three Centuries of Fundamental Change." *The Journal of Economic Perspectives* 17:4, pp. 167-190, Fall.

Lecaillon, Jacques, Felix Paukert, Christian Morrisson, and Dimitri Germidis (1984). *Income Distribution and Economic Development: An Analytical Survey*. Geneva, Switzerland: International Labour Office.

Lewis, Arthur (1954). "Development with Unlimited Supplies of Labor." *The Manchester School of Economic and Social Studies* 22:2, pp. 139-91, May.

Lindert, Peter H. (2003) *Growing Public: Social Spending and Economic Growth since the Eighteenth Century*, Cambridge: Cambridge University Press

Lindert, Peter H. (2000), Three Centuries of Inequality in Britain and America, In A. Atkinson and F. Bourguignon, eds., *Handbook of Income Distribution*, pp. 167-216. Amsterdam: North-Holland.

List, John A., and Craig A. Gallet (1999). "The Kuznets Curve: What Happens After the Inverted-U?" *Review of Development Economics* 3:2, pp. 200-206, June.

Lundberg, Mattias, and Lyn Squire (2003). "The Simultaneous Evolution of Growth and Inequality." *Economic Journal* 113:487, pp. 326-344, April.

Lybeck, J.A. and M. Henrekson (1988) Explaining the Growth of Government. Amsterdam: North-Holland.

Meltzer, Allan and Scott Richard (1981), "A Rational Theory of the Size of Government," *Journal of Political Economy*, October, vol.89, pp. 914-27.

Moav, Omer (2002). “Income Distribution and Macroeconomics: The Persistence of Inequality in a Convex Technology Framework.” *Economics Letters* 75:2, pp.187-192, April.

Morrisson, Christian (2000). “Historical Perspectives on Income Distribution: the Case of Europe.” In A. Atkinson and F. Bourguignon, eds., *Handbook of Income Distribution*, pp. 217-260. Amsterdam: North-Holland.

Mueller, Dennis C. (2003). *Public Choice III*. Cambridge: Cambridge University Press

Oshima, H.T. (1994), The Impact of Technological Transformation on Historical Trends in Income Distribution of Asia and the West, *Developing Economies*, 32, pp. 237-255

Papanek, Gustav, and Oldrich Kyn (1986). “The Effect on Income Distribution of Development, the Growth Rate and Economic Strategy.” *Journal of Development Economics* 23:1, pp. 55–65, September.

Paukert, Felix (1973). “Income Distribution at Different Levels of Development: A Survey of Evidence.” *International Labor Review* 108:2-3, pp. 97–125, August–September.

Peden, Edgar A. (1991) “Productivity in the United States and its Relationship to Government Activity: An Analysis of 57 years, 1929-86,” *Public Choice*, 86, pp. 153-73, December

Perotti, Roberto (1996), "Growth, Income Distribution and Democracy: What the Data Say," *Journal of Economic Growth*, vol. 1 (June), p.149-187.

Piketty, Thomas (1997). “The Dynamics of the Wealth Distribution and the Interest Rate with Credit Rationing.” *The Review of Economic Studies* 64:2, pp. 173-189, April.

Psacharopoulos, George (1994), Returns to Investment in Education: A Global Update, *World Development*, 22:9, pp. 1325-1343

Psacharopoulos, George and Harry Patrinos (2002), Returns to Investment in Education; A Further Update, World Bank Policy Research Working Paper 2881, The World Bank, Washington, DC

Ram, Rati (1988), Economic Development and Inequality; Further Evidence on the U-curve Hypothesis, *World Development*, 16: pp. 1371-1376

Rauch, James E. (1993). “Productivity Gains From Geographic Concentration of Human Capital: Evidence From the Cities.” *Journal of Urban Economics* 34:3, pp. 380-400, November.

Ravallion, Martin, and Shaohua Chen (1997). "What Can New Survey Data Tell Us about Recent Changes in Distribution and Poverty?" *World Bank Economic Review* 11:2, pp. 357-82.

Ravallion, Martin, and Gaurav Datt (2002). "Why Has Economic Growth Be More Pro-Poor in Some States of India Than in Others?", *Journal of Development Economics*, 68(2), pp. 381-400

Ravallion, Martin (2001), Growth, inequality and poverty : looking beyond averages, World Development, 29(11), pp. 1803-15

Robinson, Sherman (1976). "A Note on the U Hypothesis Relating Inequality and Economic Development." *American Economic Review* 66:3, pp. 437-440, June.

Rodrik, Dani (1998), "Why Do More Open Economies Have Bigger Governments?" Journal of Political Economy, 106(5), October.

Romer, Paul (1986). "Increasing Returns and Long-Run Growth." *Journal of Political Economy* 94:5, pp. 1002-37, October.

Rosenzweig, Mark (1990), "Population Growth and Human Capital Investments: Theory and Evidence," *Journal of Political Economy*, vol. 98: S12-S70

Rosenzweig, Mark and Oded Stark, eds. (1997), *Handbook of Population and Family Economics*, 2 volumes, North-Holland, Amsterdam

Rubinstein, Yona, and Daniel Tsiddon (1998), Coping with Technological Progress: the Role of Ability in Making Inequality Persistent, Processed, Tel Aviv University, Tel-Aviv

Saint-Paul, Gilles (1992), "Technological choice, financial markets and economic development,"European Economic Review, vol. 36(4), pp. 763-781

Schlicht, Ekkehart (1975). "A Neoclassical Theory of Wealth Distribution." *Jahrbücher für Nationalökonomie und Statistik* 189:1/2, pp. 78-96.

Sen, Amartya (1983), Resources, Values and Development, Blackwell, Oxford

Solow, Robert M. (1956). "A Contribution to the Theory of Economic Growth." *The Quarterly Journal of Economics* 70:1, pp. 65-94, February.

Stiglitz, Joseph (1969). "Distribution of Income and Wealth Among Individuals." *Econometrica* 37:3, pp. 382-397, August.

Syrquin, Moshe (1994). "Structural Transformation and the New Growth Theory." In L. L. Pasinetti and R. M. Solow, eds., *Economic Growth and the Structure of Long-Term Development*. New York: St.Martin's Press.

Terrell, Katherine (1992), Female-male earnings differentials and occupational structure, International Labour Review, 131:4-5, p. 387-404

Tinbergen, Jan (1975). *Income Difference: Recent Research*. Amsterdam: North-Holland.

Tressel, Thierry (2003), "Dual financial systems and inequalities in economic development, *Journal of Economic Growth*, vol. 8, pp. 223-257

Thomas, Vinod, Yan Wang, and Xibo Fan (2000). "Measuring Education Inequality: Gini Coefficients of Education." World Bank Working Paper 2525, December.

Townsend, Robert M., and Kenichi Ueda (2003). "Financial Deepening, Inequality, and Growth: A Model-Based Quantitative Evaluation." International Monetary Fund Seminar Series 40, pp. 1-64, March.

Tsaklogou, P. (1988), Development and Inequality Revisited, *Applied Economics*, 20: 509-531

Violante, Giovanni (2002), Technological Innovation, Skill Transferability and the Rise in Residual Inequality, *Quarterly Journal of Economics*, 117(1), pp. 297-338

Wagner, Adolf (1883). *Finanzwissenschaft*. Extracts in: Richard A. Musgrave and Alan R. Peacock eds. *Classics in the Theory of Public Finance*. London: Macmillan, 1958, pp. 1-8.

Welch, Finis (2000), Growth in Women's Relative Wages and in Inequality Among Men: One Phenomenon or Two? *American Economic Review*, 90:2, pp. 444-449

World Bank (1990), *World Development Report 1990. Poverty*. Oxford University Press, published for the World Bank.

World Bank (2000), *Attacking Poverty, World Development Report 2000/2001*. Oxford University Press, published for the World Bank.

World Bank (2001). *Engendering Development*, Policy Research Report, Oxford University Press, published for the World Bank.

World Bank (2003). *Statistical Information Management & Analysis (SIMA)*. Online database.

Zhang, Xiabo, Michael Johnson, Danielle Resnick and Sherman Robinson (2004). "Cross-Country Typologies and Development Strategies to End Hunger in Africa," International Food Policy Research Institute DSDG Discussion Paper No.8, June

Table 1. Estimated growth elasticity of selected economic and socio-demographic indicators

	Simple regression on country means		Pooling		Fixed effects		Decadal differences	
Dependent variable	**1970s**	**1990s**	**No other variable**	**+ Non-linear Trend**	**No other variable**	**+ Non-linear Trend**		
Public expenditures (% GDP)	0.47	0.63	0.50	0.58	0.07	0.25	0.27	
T-statistic	*(0.86)*	*(5.19)***	*(12.22)***	*(15.58)***	*(1.74)*	*(4.12)***	*(1.26)*	
# observations	*105*	*131*	*2546*	*2546*	*2546*	*2546*	*202*	
Agricultural/non-agricultural productivity differential	na	0.01	0.01	0.01	0.005	0.01	0.09	
T-statistic	*na*	*(2.38)**	*(6.71)***	*(6.49)***	*(2.92)***	*(3.89)***	*(2.23)**	
# observations	*na*	*91*	*1323*	*1323*	*1323*	*1323*	*53*	
Urbanization rate (Cox)	0.49	0.35	0.33	0.34	0.15	0.11	0.11	
T-statistic	*(5.35)***	*(3.23)***	*(14.38)***	*(15.08)***	*(30.33)***	*(17.79)***	*(5.33)***	
# observations	*118*	*166*	*4008*	*4008*	*4008*	*4008*	*265*	
Literacy rate among adults (Cox)	3.57	2.79	3.67	3.21	1.74	1.06	0.58	
T-statistic	*(1.6)*	*(2.53)**	*(11.71)***	*(10.58)***	*(18.48)***	*(8.94)***	*(1.32)*	
# observations	*90*	*126*	*3056*	*3056*	*3056*	*3056*	*200*	
Average schooling year in total population over 25	0.76	0.30	0.33	0.33	0.15	0.02	0.02	
T-statistic	*(8.21)***	*(17.60)***	*(28.00)***	*(27.69)***	*(17.86)***	*(2.01)**	*(1.51)*	
# observations	*93*	*99*	*575*	*575*	*575*	*575*	*189*	
Life expectancy (years)	2.94	1.01	0.97	0.94	0.30	0.05	0.06	
T-statistic	*(11.36)***	*(12.00)***	*(29.77)***	*(28.15)***	*(20.17)***	*(2.66)***	*(1.79)*	
# observations	*114*	*167*	*2219*	*2219*	*2219*	*2219*	*260*	
Fertility (number of children)	-0.49	-0.15	-0.15	-0.12	-0.04	0.07	0.06	
T-statistic	*(7.60)***	*(11.49)***	*(28.94)***	*(25.05)***	*(12.23)***	*(26.99)***	*(8.46)***	
# observations	*119*	*167*	*2526*	*2526*	*2526*	*2526*	*265*	
Male/female differential in adult literacy (%)	-1.49	-0.94	-1.02	-0.99	-0.56	-0.20	-0.19	
T-statistic	*(1.88)*	*(5.80)***	*(25.60)***	*(23.35)***	*(29.60)***	*(8.96)***	*(3.28)***	
# observations	*90*	*126*	*3056*	*3056*	*3056*	*3056*	*200*	
Male/female differential in life expectancy (year)	-0.45	-0.13	-0.12	-0.11	0.01	0.02	0.01	
T-statistic	*(8.06)***	*(7.81)***	*(21.16)***	*(17.92)***	*(3.51)***	*(3.12)***	*(0.9)*	
# observations	*114*	*167*	*2219*	*2219*	*2219*	*2219*	*260*	

	(y)	(y²)	(y)	(y²)	(y)	(y²)	(y)	(y²)	(y)	(y²)	(y)	(y²)	(y)	(y²)
Income inequality (Gini)	-1.60		-0.27		-0.51		-0.50		0.07		0.11		0.04	
T-statistic	*(4.40)***		*(2.69)***		*(8.08)***		*(7.24)***		*(1.41)*		*(1.14)*		*(0.31)*	
# observations	*52*		*54*		*389*		*389*		*389*		*389*		*92*	
Income inequality (Gini)	2.37	-0.43	-0.03	-0.01	-1.11	0.03	-1.06	0.03	-0.16	0.01	-0.15	0.01	-0.43	0.01
T-statistic	*(1.21)*	*(2.05)**	*(0.06)*	*(0.43)*	*(5.31)***	*(3.28)***	*(4.52)***	*(2.64)***	*(0.81)*	*(1.22)*	*(0.44)*	*(0.78)*	*(1.36)*	*(1.83)*
# observations	*52*	*52*	*54*	*54*	*389*	*389*	*389*	*389*	*389*	*389*	*389*	*389*	*92*	*92*
Female labor force participation (%)	-0.83		0.12		0.19		0.09		0.53		0.33		0.34	
T-statistic	*(2.17)**		*(1.23)*		*(8.42)***		*(4.01)***		*(52.33)***		*(27.55)***		*(4.74)***	
# observations	*113*		*156*		*3811*		*3811*		*3811*		*3811*		*251*	
Female labor force participation (%)	-5.25	0.49	-0.65	0.03	-0.20	0.02	-0.29	0.02	1.12	-0.02	0.87	-0.01	0.86	-0.01
T-statistic	*(2.79)***	*(2.19)**	*(1.35)*	*(1.51)*	*(1.58)*	*(2.77)***	*(2.44)**	*(2.85)***	*(50.98)***	*(29.48)***	*(31.42)***	*(21.32)***	*(5.66)***	*(4.05)***
# observations	*113*	*113*	*156*	*156*	*3811*	*3811*	*3811*	*3811*	*3811*	*3811*	*3811*	*3811*	*251*	*251*

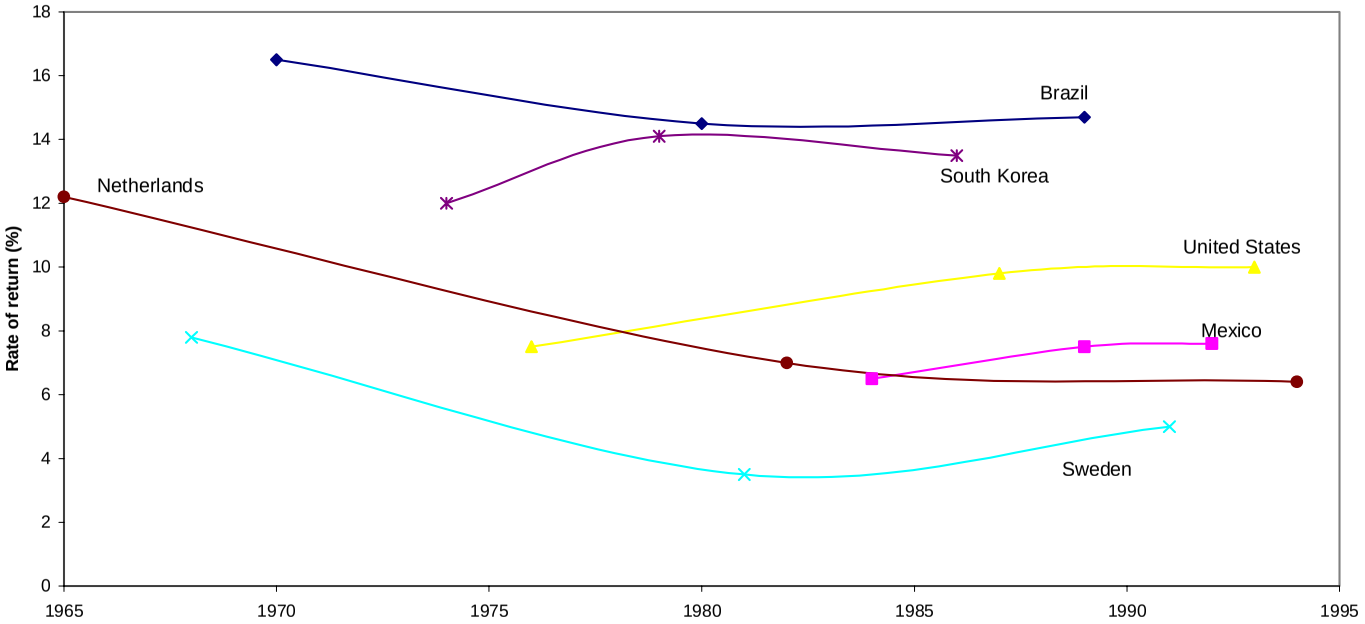
Source : author's calculation on WDI data
Robust T-statistics in brackets
Cox transformation of variable x is x/(1-x).
* significant at 5%; ** significant at 1%

Table 2. Earning differentials by educational levels (%)

	Primary vs no schooling	Secondary vs. Primary	Tertiary vs. Secondary	Average earning differential by year schooling
Low income countries (GDP per capita less than $755)	25.8	19.9	26	10.9
Middle-income countries (to $9265)	27.4	18	19.3	10.7
High income countries (GDP per capita more than $9265)	n.a.	12.2	12.4	7.4

Source : Psacharopoulos and Patrinos (2002)

Figure 1. Evolution of rate of return to education (years of schooling)



Source : Based on Psacharopoulos and Patrinos (2002)

Economic Growth and the Environment: A Review of Theory and Empirics.

William A. Brock
Vilas Professor of Economics, Department of Economics, University of Wisconsin - Madison
E-mail: wbrock@ssc.wisc.edu

M. Scott Taylor
Professor of Economics, Department of Economics, University of Wisconsin-Madison
and
Faculty Research Associate, National Bureau of Economics Research, Cambridge, MA.
E-mail: staylor@ssc.wisc.edu
web site: www.ssc.wisc.edu/~staylor

September 14, 2004

1 Introduction

The relationship between economic growth and the environment is, and may always remain, controversial. Some see the emergence of new pollution problems, the lack of success in dealing with global warming and the still rising population in the Third World as proof positive that humans are a short-sighted and rapacious species. Others however see the glass as half full. They note the tremendous progress made in providing urban sanitation, improvements in air quality in major cities and marvel at the continuing improvements in the human condition made possible by technological advance. The first group focuses on the remaining and often serious environmental problems of the day; the second on the long, but sometimes erratic, history of improvement in living standards.

These views are not necessarily inconsistent and growth theory offers us the tools needed to explore the link between environmental problems of today and the likelihood of their improvement tomorrow. It allows us to clarify these conflicting views by use of theory, and when differences still remain, to create useful empirical tests that quantify relative magnitudes.

For many years, the limited natural resource base of the planet was viewed as *the* source of limits to growth. This was, for example the focus of the original and subsequent "Limits to Growth" monograph and the efforts by economists refuting its conclusions.[1] Recently however it has become clear that limits to growth may not only arise from nature's finite source of raw materials, but instead from nature's limited ability to act as a sink for human wastes. It is perhaps natural to think first of the environment as a source of raw materials, oil and valuable minerals. This interpretation of nature's service to mankind led to a large and still growing theoretical literature on the limits to growth created by natural resource scarcity. Empirically it led to studies of the drag limited natural resources may have on growth, and a related examination of long run trends in resource prices.[2]

Nature's other role - its role as a sink for unwanted by-products of economic activity - has typically been given less attention. As a sink, nature dissipates harmful air, water and solid pollutants, is the final resting

[1] See Nordhaus (1992) for the latest refutation.
[2] For work on resource price trends see most importantly Barnett and Morse (1963) and Slade (1982).

place for millions of tons of garbage, and is the unfortunate repository for many toxic chemicals. When the environment's ability to dissipate or absorb wastes is exceeded, environmental quality falls and the policy response to this reduction in quality may in turn limit growth. Growth may be limited because reductions in environmental quality call forth more intensive clean up or abatement efforts that lower the return to investment, or more apocalyptically, growth may be limited when humans do such damage to the ecosystem that it deteriorates beyond repair and settles on a new lower, less productive steady state.[3]

This link between growth and the environment has of course received much more attention recently because of the rapidly expanding empirical literature on the relationship between per capita income and pollution. This literature, known as the Environmental Kuznets Curve (EKC) literature, has been enormously influential. So to a certain extent, the tables have now turned: there is far less concern over the ultimate exhaustion of oil or magnesium, and far more concern over air quality, global warming, and the emissions of industrial production.

The economics literature examining the link between growth and the environment is huge; it covers, in principle, much of the theory of natural resource extraction, a significant body of theory in the 1960s and 1970s on resource depletion and growth; a large literature in the 1990s investigating the implications of endogenous growth theories; and a new and still growing literature created in the last decade examining the relationship between pollution and national income levels. Every review has to make difficult choices about exclusion and we make ours on the basis of novelty. There are excellent book length treatments on the depletion of renewable and non-renewable resources, and several reviews of endogenous growth theory's contributions already exist.[4] This leaves us to focus on the relatively new theoretical literature linking environmental quality to income levels. For the most part, we discuss the link between industrial pollution and growth, but also show how this most recent work is related to earlier contributions on exhaustible resources and growth.

While no review can settle the perennial debate over the limits to growth, this review hopes to play a positive role in moving the literature forward by identifying important unresolved theoretical questions, reporting on the results of recent empirical work, and providing an integrative assessment of where we stand today.[5] To do so, we focus on three questions. These are: (1) what is the relationship between economic growth and the environment?; (2) how can we escape the limits to growth imposed by environmental constraints?; and (3), where should future research focus its efforts?

To answer these questions we start by introducing definitions and providing a preliminary result linking the environment and growth. We define the scale, composition and technique effects of growth on the environment, and then use these definitions to prove a useful but negative result on the limits to growth. We show that changes in the composition of national output – as occur when the economy specializes in relatively less pollution intensive services or relatively less natural intensive industries- can at best delay the impact of binding environmental constraints. In the long run, emission intensities must fall towards zero if growth is to be sustainable.

In many models this constraint is met through the substitution of clean inputs for dirty ones, in others via increased abatement, and in still others through some combination of technological progress and the other channels. This result is helpful to us because it allows us to distinguish between empirical regularities that are consistent with a short run growth and environment relationship (along a transition path) from those consistent with the long run relationship (along a balanced growth path). It also helps us sort through the literature by focusing on how a given model can generate what we take as our definition of sustainable

[3] This branch of the literature relies on case study evidence of irreversible damage created in the past and argues that our now greater technological capabilities may portend even worse outcomes in the future. For a primarily theoretical discussion of irreversibilities and hysteresis caused by nonlinearities see the symposium edited by Dasgupta and Maler (2003). For related nonlinear theory see Dechert (2001). For case study evidence from prehistory see Brander et al. (1998).

[4] See the classic book length treatments of renewable and nonrenewable resources by Clark (1990), and Dasgupta and Heal (1979). A good introduction to the relationship between endogenous growth theory and the environment is contained in the review by Smulders (1999).

[5] Whether there are serious limits to growth is an unending controversy that reached its peak with the publication of the Limits to Growth by Meadows et al. in 1972. See the subsequent contributions by Solow (1973) followed by Meadows et al. (1991) and then Nordhaus (1992).

growth: a balanced growth path with increasing environmental quality and ongoing growth in income per capita.[6]

With our definitions and result in hand we then turn to present some stylized facts on the environment and growth. These facts concern the trend and level of various pollutants, and measures of the cost of pollution control. In many cases, the data underlying the construction of these facts is of limited quality; the time periods are sometimes insufficiently long to draw strong conclusions and the relevant magnitudes imprecise relative to their constructs in theory. Nonetheless, they are the best data we have.

Overall these data tell three stories. The first is that by many measures the environment is improving at least in developed countries. The level of emissions for regulated pollutants is falling, and the quality of air in cities is rising. The U.S. and other advanced industrial countries have seen secular improvements in the quality of their environments over the last 30 years. To a large extent cities are cleaner than in the past, emissions of health-threatening toxics are reduced, and in some cases the changes in environmental quality are quite dramatic.

The second feature of the data is that pollution control measures have been both relatively successful and relatively cheap. While there are severe difficulties in measuring the full cost of environmental compliance most methods find costs of at most 1-2% of GDP for the U.S. Comparable figures from OECD countries support this finding.[7]

The last feature of the data is that there is a tendency for the environment to at first worsen at low levels of income but then improve at higher incomes. This is the so-called Environmental Kuznets Curve. We first present raw emission data drawn from the U.S. and then briefly review the empirical literature on the Environmental Kuznets's Curve that relies on cross-country comparisons. The raw data from the US are unequivocal, while the cross-country empirical results are far less clear but generally supportive of the finding.

Having reviewed the relevant data and set out definitions we turn to a review of the theory. To do so, we develop a series of 4 simple growth and environment models. The models serve as a vehicle to introduce related theoretical work. For the most part we focus on balanced growth path predictions and eschew formal optimization taking as exogenous savings or depletion rates and sometimes investments in abatement. We do so because in many cases, these rates must be constant along any balanced growth path and hence we identify a set of feasible conditions for sustainable growth. Moreover the resulting simplicity of the models allows us to identify key features of fully developed research contributions already present in the literature. In some cases, the choice of abatement or savings matters critically to the point we are making and hence in those cases we provide optimal rules.

The 4 models were developed to highlight the different ways we can meet environmental constraints in the face of ongoing growth in per capita incomes. In the first, which we dub the Green Solow Model, emission reductions arise from exogenous technological progress in the abatement process. Although this model is very simple it provides three useful results. First, we show that even with the economy's abatement intensity fixed, the dynamics of the Solow model together with those of a standard regeneration function are sufficient to produce the Environmental Kuznet's Curve. The transition towards any sustainable growth path has environmental quality at first worsening with economic growth and then improving as we approach the balanced growth path. This is a surprising result. While numerous explanations for the EKC relationship have been put forward, this explanation is simple, novel, and quite general as it relies only on basic properties of growth functions.

Second, the Green Solow model provides a useful benchmark since this model predicts that a more strict pollution policy has no long run effect on growth. In true Solow tradition, different abatement intensities

[6] This is different from other definitions. We wanted to avoid stagnation as a sustainable growth path and hence require positive growth; but with positive income growth giving more marketable goods along the balanced growth path it seems only appropriate to require an improving environment as well since this gives us more non market goods.

[7] Aggregate compliance costs were reported in a 1990 EPA study that has apparently never been updated. (See EPA (1990) Environmental Investments: The cost of a clean environment) The earlier study predicted year 2000 compliance costs of approximately 200 billion dollars (1990 dollars), but recent EPA publications (EPA's 2004 Strategic Plan) distances themselves from this estimate and reiterates just how difficult it is to estimate compliance costs. OECD evidence can be found in the publication Pollution Abatement & Control Expenditures in OECD countries, Paris: OECD Secretariat.

create level differences in income but have no effect on the economy's growth rate along the balanced growth path. This result provides partial justification for the current practice of measuring the costs of pollution control as the sum of current private and public expenditures with no correction for the reduction in growth created. It also points out the stringent conditions needed for a stricter policy policy to cause no drag whatsoever on economic growth.

Third, the model clearly shows how technological progress in goods production has a very different environmental impact than does technological progress in abatement. Technological progress in goods production creates a scale effect that raises emissions, technological progress in abatement creates a pure technique effect driving emissions downwards. In the Green Solow model both rates are exogenous, and as such they provide especially clean examples of scale and technique effects for us to refer to later. And as we show throughout the review, the presence or absence of technological progress in abatement is key to whether we can lower emissions, support ongoing growth, and provide reasonable predictions for the costs of pollution control.

The second model, which we dub the Stokey Alternative, was inspired by Nancy Stokey's (1998) influential paper on the limits to growth. Here we present a simplified version to highlight the role abatement can play in improving the environment over time. The model we present focuses on balanced growth paths and not the transition paths as emphasized by Stokey, but nevertheless it contains two results worthy of note. The first is simply the observation that once we model abatement as an economic activity that uses scarce resources, increases in the intensity of abatement that are needed to keep pollution in check will have a drag on economic growth. Rising abatement creates a technique effect by lowering emissions per unit output, but also lowers pollution by lowering the growth rate of output.

By rewriting the model along the lines of Copeland and Taylor (1994) so that pollution emissions appear as if they are a factor of production, it is now relatively simple to conduct growth drag exercises for the cost of pollution control in much the same way that others have examined the growth drag of natural resource depletion.[8] By doing so, the model makes clear the limits to growth brought about by environmental policy.

The second feature we focus on is the model's prediction concerning the intensity of abatement. In models with falling pollution levels, neoclassical assumptions on abatement, and no abatement specific technological progress, the intensity of abatement must rise continuously through time. For example, in Stokey's analysis the share of "potential output" allocated to abatement approaches one in the limit. Since this share represents pollution abatement costs relative to the value of aggregate economic activity, models that rely on abatement alone tend to generate counterfactual predictions for abatement costs. This is true even though ongoing economic growth is fueled by technological progress, and hence this result reinforces our earlier remarks about the importance of technological progress in abatement.

Our third model links the source and sink roles of nature by assuming energy use both draws down exhaustible resource stocks and creates pollution emissions that lower environmental quality. This "source and sink" formulation allows us to examine how changes in the energy intensity of production help meet environmental constraints. In this model, the intensity of abatement is taken as constant and there is no technological progress in abatement. Instead the economy lowers its emissions to output ratio over time by adopting an ever cleaner mix of production methods. As such the model focuses on the role of composition effects in meeting environmental constraints. We show that the economy is able to grow while reducing pollution because of continuous changes in the composition of its inputs, but this form of "abatement" has costs. Growth is slowed as less and less of the natural resource can be used in production.

This "source and sink" formulation is important in linking the earlier 1970s and 1980s literature focusing on growth and resource exhaustion with the newer 1990s literature focusing on the link between economic growth and environmental quality. We show that the finiteness of natural resources implies a constraint on per capita income growth that is worsened with higher population growth rates. This constraint is relaxed if the rate of natural resource use is slower as this implies reproducible factors have less of a burden in keeping growth positive. But sustainability also requires falling emissions, and this constraint is most easily met if the economy makes a rapid transition away from natural resource inputs as this reduces the energy and pollution intensity of output.

[8] See for example Nordhaus (1992).

Putting the constraints from the source and sink side together, we show there exist parameter values for which the twin goals of positive ongoing growth and falling emission levels are no longer compatible. This is not a doomsday prediction. Together with our previous analysis it suggests that abatement or composition shifts alone are unlikely to be responsible for the stylized facts. Technological progress directly targeted to lowering abatement costs (i.e. induced innovation) must be playing a key role in determining growth and environment outcomes. Therefore, in the remainder of the paper we turn to a model where technological progress in abatement is set in motion by the onset of active regulation and works to generate sustainable growth paths.

To highlight the importance of technological progress in abatement our final model draws on the analysis of Brock and Taylor (2003) by adopting their Kindergarten Rule model. While the previous models were useful vehicles to discuss the literature and describe possibilities, they were necessarily incomplete because they eschew formal optimization. Optimizing behavior is however important in discussions of the magnitude of drag created by pollution policy, and also important in discussions concerning the timing or onset of active regulation. The Kindergarten model provides two contributions to our discussion.

First, it shows how technological progress in abatement can hold compliance costs down in the face of ongoing growth. In contrast to the Green Solow model, there are ongoing growth drag costs from regulation, but as long as abatement is productive it is possible to generate sustainable growth without skyrocketing compliance costs. By highlighting the important role for progress in abatement, the model points out the need to make endogenous the direction of technological progress as well as its rate.

Second, the model generates a first worsening and then improving environment much like that in Stokey (1998). In contrast however to the methods employed in the empirical EKC literature, we show that the path for income and pollution will differ systematically across countries. This systematic difference leads to the model's Environmental Catch-up Hypothesis relating income and pollution paths to countries initial income levels. Poor countries experience the greatest environmental degradation at their peak, but once regulation begins environmental quality across both Rich and Poor converges. Despite this, at any given income level an initially Poor country has worse environmental quality than an initially Rich country. Moreover, since both Rich and Poor economies start with pristine environments, the qualities of their environments at first diverge and then converge over time. In addition to this cross-country prediction, the model also links specific features of the income and pollution profile to characteristics of individual pollutants such as their permanence in the environment, their toxicity, and their instantaneous disutility. Together these predictions suggest a different empirical methodology than that currently employed, and expand the scope for empirical work in this area considerably.

The final section of our review is a summary of the main lessons we have drawn from the literature, offers suggestions for future research and briefly discusses some of the most important topics that we did not discuss elsewhere in the review.

2 Preliminaries

2.1 Scale, Composition and Technique

We start with some algebra linking emissions of a given pollutant to a measure of economic activity, its composition and the cleanliness of production techniques. By doing so we illustrate that any growth model that predicts both rising incomes and falling pollution levels has to work on lowering pollution emissions via one of three channels. Consider a given pollutant and let E denote the sum total of this pollutant's emissions arising from production across the economy's n industries.[9] Let a_i denote the pounds of emissions per dollar of output produced in industry i, s_i denote the value share of industry i in national output, and

[9] The pollutant could instead be produced via consumption. In that case we adopt weights reflecting industry i's share in final demand. This has little impact on our results here, but would have some relevance in an open economy setting.

Y national output. Then by definition total emissions E are given by:

$$E = \sum_{i=1}^{n} a_i s_i Y \text{ where } \sum_{i=1}^{n} s_i = 1 \tag{1}$$

Since this is a definition we can differentiate both sides with respect to time to find:

$$\hat{E} = \sum_{i=1}^{n} \pi_i \left[\hat{a}_i + \hat{s}_i\right] + \hat{Y} \text{ where } \pi_i = \frac{E_i}{E} \tag{2}$$

where a $\hat{}$ over x indicates $[dx/dt]/x$. Changes in aggregate emissions can arise from three sources that we define to be the Scale, Composition and Technique effects.[10]

To start, note that holding constant the cleanliness of production techniques and the composition of final output (i.e. holding both $\hat{a}_i = 0$ and $\hat{s}_i = 0$ for all i) emissions rise or fall in proportion to the scale of economic activity as measured by real GDP or Y. This is the scale effect of growth and unless it is offset by other changes, emissions rise lock step with increases in real output.

Alternatively, we can hold both the scale of real output and the techniques of production constant to examine the impact of changes in the composition of output. To do so, in (2) we set $\hat{Y} = 0$ and $\hat{a}_i = 0$ for all i as this isolates the pure composition effect on pollution emissions.

$$\hat{E} = \sum_{i=1}^{n} \pi_i \hat{s}_i \tag{3}$$

Emissions fall via the pure composition effect if an economy moves towards producing a set of goods that are cleaner on average than the set they produced before. To see why this is true, note that the change in value shares across all n industries must sum to zero; i.e. $\sum_{i=1}^{n} ds_i = 0$. Now using this result in (3) we obtain the change in emissions arising from a pure composition effect as:

$$\hat{E} = \sum_{i=1}^{n} \hat{s}_i \left[\pi_i - s_i\right] \tag{4}$$

Given our definitions $\pi_i - s_i > 0$ if and only if $E_i/p_i y_i > E/Y$. In words, the element $\pi_i - s_i$ is positive if and only if industry i's emissions per dollar of output is greater than the national average. Define a dirty industry as one whose emissions per dollar of output exceed the economy wide average E/Y; define a clean industry as one where emissions per dollar of output are less than the economy average. Then equation (4) holds that aggregate emissions fall from the pure composition effect whenever the composition of output changes toward a more heavy reliance on clean industries and rises otherwise.

Finally, emissions can fall when the techniques of production become cleaner even though output and its composition remain constant. To isolate this technique effect, we set $\hat{Y} = 0$ and $\hat{s}_i = 0$ for all i to find that emissions fall if emissions per unit output fall for all activities. In this case we find:

$$\hat{E} = \sum_{i=1}^{n} \pi_i \hat{a}_i \tag{5}$$

and hence if techniques are getting cleaner, emissions per unit of output fall, and overall emissions fall from this pure technique effect.

When the environment is modeled as a sink for human wastes it is often assumed that emissions together with natural regeneration determine environmental quality. When the environment adjusts relatively slowly to changes in the pollution level, natural regeneration can play an important role in determining environmental quality. A typical and very useful specification assumes the environment dissipates pollutants at an

[10] See Copeland and Taylor (1994) for model based definitions of these effects in a static setting. This terminology was popularized by Grossman and Krueger's (1993) NAFTA study.

exponential rate. Let X denote the pollution stock (an inverse measure of environmental quality) and let the pristine level be given by $X = 0$. Then since the flow of emissions per unit time is E, the evolution of the pollution stock is given by:

$$\dot{X} = E - \eta X \text{ where } \eta > 0 \tag{6}$$

This formulation is convenient because it is generally assumed that X must be bounded for human life to exist and hence (6) yields a simple negative linear relationship between the steady state flow of pollution E, and the pollution stock X. A bound on X then implies a similar bound on steady state emissions, E.[11] Moreover, given the linear relationship any scale, composition or technique effect on emissions is translated directly into impacts on X.

One cost of (6) is that the percentage rate of natural regeneration is independent of the state of the environment. A common modification is to assume the rate of natural regeneration rises as X gets further and further from its pristine level. Letting $\eta = \eta(X)$, we can introduce this possibility by writing the evolution of X as:

$$\dot{X} = E - \eta(X) X \text{ where } \eta'(x) > 0 \tag{7}$$

$\eta(X)$ is often assumed to be linear.

An alternative and equally valid interpretation of (2) is that E is the instantaneous flow of natural resources used in production. Under this interpretation, equation (1) gives us an economy wide factor demand for this natural resource evaluated at the equilibrium level of use given by E. For example, the demand for oil equals the sum of demand arising from all sectors of the economy. In this interpretation a_i are barrels of oil used per unit of output in industry i, s_i is the value share of industry i in national output, and Y is again national output.

For example, if the flow of resources extracted is falling at some constant rate over time while real output is rising, then we know that some combination of changes in energy efficiency per unit of output (a technique effect) and changes in the output mix to less energy intensive goods (a composition effect) must be carrying the burden of adjustment. Changes in resource use over time can then be linked to the relative strength of scale, composition and technique effects. To complete the translation let the current stock of natural resources S be given by our initial endowment K less the sum over time of extraction by humans, E. If this resource has a zero regeneration rate we obtain the standard equation governing stock depletion in exhaustible resources:

$$\dot{S} = -E \tag{8}$$

Alternatively, we can leave open the possibility of regeneration. Making the needed changes to (8) gives us the standard accumulation equation for a renewable resource such as a forest or fishery when growth is stock dependent.

$$\dot{S} = \eta(S) S - E \tag{9}$$

And again if $\eta(S)$ is linear we obtain the familiar logistic growth for a naturally regenerating resource.[12]

Although (1) is a definition it implicitly contains an assumption on how economic growth and the environment interact. Note that the value shares sum to 1 and $a_i(t) \geq 0$ for all i and t. Assume that $a_i(t) > 0$ for all i and t. This assumption turns out to be an important, because if some activities are perfectly clean, or approach perfectly clean activities in the limit, then it is possible for composition effects alone to hold pollution in check despite ongoing growth. Conversely, if all economic activities must pollute even a small

[11] Along a balanced growth path the time rate of change of X must equal that of E. To see this divide both sides of (6) by X and note that a constant rate of change in X requires the ratio E/X to be constant.

[12] It should be noted however that different assumptions on $\eta(S)$ can lead to drastically different conclusions when they lead to growth functions with what biologists call critical depensation [See Clark (1990) for a formal definition and discussion]. Critical depensation refers to a property of the natural growth function such that at some minimum S, natural growth becomes negative. Natural growth can turn negative because of predator prey interactions across species, or because the species has a minimum viable population. Introducing thresholds and critical depensation into either (9) or (7) can alter results considerably. Unfortunately little is known about the extent of non-convexities of this type empirically. For theoretical work examining their impact see the symposium edited by Dasgupta and Maler (2003). Scheffer and Carpenter (2003) document some examples of catastrophic regime shifts in ecosystems.

amount, then environmental quality can only rise in the long run via continuous changes in the techniques of production and these may run into diminishing returns.

It is not helpful here to enter into philosophical discussions over the definition of pollution or the likelihood of today's unwanted outputs becoming tomorrow's valuable inputs. Instead we just note that all production involves the *transformation* of one set of materials into another and that this transformation requires work. All work requires energy and energy is always wasted in work effort. Therefore some unintentional by products of production are always produced and we most often call these by products pollution. Since this is a statement of belief and not a rigorous proof, we note this as an assumption.

Assumption 1. Pollution is a by-product of all production:

$$for\ all\ i,\ t \geq 0,\ lim\ inf\ \{a_i(t)\} > 0 \tag{10}$$

this implies that there exists for each i, a strictly positive $\varepsilon > 0$ such that $a_i(t) > \varepsilon$. With Assumption 1 in hand, it is now possible to show that composition effects are at best a transitory method to lower pollution emissions. Let us explain in detail why this conclusion holds. Suppose there is a bound, $B > 0$ such that if $E(t)$ exceeds B, human life cannot exist. Then if $Y(t)$ goes to infinity as t goes to infinity, (10) implies:

$$E(t) \leq B \Rightarrow \sum_{i=1}^{n} a_i(t)\, s_i(t) \leq \frac{B}{Y(t)}\ for\ all\ t \geq 0 \tag{11}$$

Thus we must have

$$\overset{*}{a}_i(t) := \min\{a_i(t)\} \leq \frac{B}{Y(t)} \to 0\ as\ t \to \infty \tag{12}$$

But (12) contradicts Assumption 1. Hence if we are to have bounded emissions with growing $Y(t)$, we must have the cleanest industry emission rate $\overset{*}{a}_i(t)$ going to zero. Therefore, falling pollution levels and rising incomes are only possible if there are continual reductions in emissions per unit output and zero emission technologies are possible, at least in the limit.

3 Stylized Facts on Sources and Sinks

We present three stylized facts drawn from post WW II historical record. We present data on pollution emissions and environmental control costs and leave the discussion of energy prices to later sections. Since data is typically only available for pollutants that are presently under active regulation we discuss the US record with regard to its six so-called criteria air pollutants, but amend these with international sources where possible. These are: sulfur dioxide, nitrogen oxides, carbon monoxide, lead, large particulates and volatile organic compounds.[13] With the exception of lead, these air pollutants all typically classified as irritants and so we also briefly discuss the US history of regulation of long-lived and potentially harmful chemical products. For the most part we present data on emissions rather than concentrations because data on emissions covers a much longer time period and is unaffected by industry location and zoning regulation. On the other hand, the longest time spans of data (from 1940 onwards) reflect some changes in collection and estimation methods.[14] Nevertheless, this data is the best we have available and where possible we direct the reader to concentration data and related empirical work. In addition we present data on industry pollution abatement costs from Vogan (1996), although these are only available for the 1972-1994 period.

[13] The long series of historical data presented in the figures is taken from the EPA's 1998 report National Pollution Emission Trends, available at http://www.epa.gov/ttn/chief/trends/trends98.

[14] As methods of estimation improve new categories of emissions are included and some revision occurs as well. For example, prior to 1985 the PM10 data excluded fugitive dust sources and other miscellaneous emissions, so these are eliminated from the time series graphed in Figure 7. As well revision occurs. A close look at the 2001 Trends report shows that emissions reported for our pollutants during the 1970s and 1980s does not exactly match the figures given in the 1998 report. We use the 1998 figures rather than those from 2001 since the 2001 report only contains estimates to 1970, and we import the EPA's graphics directly into our figures because we cannot match them precisely from the raw data.

We start by presenting in Figure 1 emissions per dollar of GDP for all pollutants except lead. Lead is excluded since data is only available over a much shorter period. As shown, emissions per unit of output for sulfur, nitrogen oxides, particulates, volatile organic compounds, and carbon monoxide all fall over the 1940-1998 period. For ease of comparison emission intensities were normalized to 100 in 1940 and the figure adopts a log scale. PM10 fell by approximately 98%, sulfur, volatile organic compounds and carbon monoxide fell by perhaps 88%, and nitrogen oxides fell by perhaps 60%. Somewhat surprisingly, it is also apparent that if we exclude the years of WWII at the start of the data, the rate of reduction for each pollutant appears to be roughly constant over time.

Although there is a tendency to see good news in falling emission intensities, there are good reasons for not doing so. One reason is simply that real economic activity increased by a factor of 8.6 over this period and this masks the fact that emissions of many of these pollutants rose during this period. The second is that this measure – like that for aggregate emissions – has very little if any welfare significance. Since our measure is physical tons of emissions added up over all sources, it necessarily ignores the fact that some tons of emissions create greater marginal damage than others.[15]

Our second stylized fact is presented in Figure 2. In it we plot business expenditures on pollution abatement costs per dollar of GDP over the period 1972-1994. These twenty-two years are the only significant time period where data is available.[16] As shown, pollution abatement costs as a fraction of GDP rise quite rapidly until 1980 and then remain relatively constant. As a fraction of overall output, these costs are relatively small. Alternatively, if we consider pollution abatement costs specifically directed to the six criteria air pollutants and scale this by real US output, the ratio is then incredibly small – approximately one half of one percent of GDP - and has remained so for over twenty years (See Vogan (1996)).[17]

Data from other countries supports the general conclusion that pollution abatement costs are a small fraction of GDP and show perhaps a slight upward trend. For example, total expenditures by both government and business in France rose from 1.2% of GDP in 1990 to 1.6% in 2000. Over the 1991-1999 period, these same expenditures in Germany rose from 1.4% of GDP to 1.6%. Austria and the Netherlands show somewhat higher expenditures on the order of 2.1% and 1.6% in 1990 rising to 2.6% and 2.0% in 1998. While this data is clearly fragmentary, expenditures in the order of 1-2% of GDP seem to be the norm in OECD countries, with perhaps half of this being spent by private establishments and the remainder by governments.[18]

These figures however reflect to a certain degree the changing composition of output over time and therefore understate the impact higher pollution abatement costs have had on some industries. Levinson and Taylor (2003) for example argue that since the composition of U.S. manufacturing has been shifting towards less pollution intensive industries, aggregate measures understate the true costs of pollution regulations. They construct estimates of pollution abatement costs holding the composition of industry output fixed at the 2 and 3 digit industry levels and then compare these estimates with estimates allowing the composition of output to change. In all cases, holding the composition of US output fixed in earlier periods leads to a higher estimate of industry wide abatement cost increases. As a result, the small increases in pollution abatement costs shown in the aggregate data are at least partially due to the U.S. shedding some of its dirtiest industries over time.

Our third and final fact is presented in Figures 3 to 8. These figures show a general tendency for emissions to at first rise and then fall over time. Note that the falling emissions/intensities reported in Figure 1 are necessary but not sufficient for this result. This pattern in the data is visible for all pollutants except nitrogen

[15] In contrast, a quality-adjusted measure of emissions would add up the various components weighing them by their marginal damage; or a quality-adjusted measure of aggregate concentrations in a metropolitan area would weigh concentrations in each location by the marginal damage of concentrations at point (urban, industrial, suburban, etc.).

[16] In 1999 the PACE survey was run again this time as a pilot project. Using the 1999 survey we find the ratio of PACE to GDP of approximately 1.9% which is very much in line with Figure 2. This 1999 survey is different in some respects from earlier ones. For details see the Survey of Pollution Abatement Costs and Expenditures, U.S. Census Bureau 1999 available at www.census.gov/econ/overview/mu1100.html

[17] These figures are also similar to those presented in the review of pollution abatement costs in Jaffe et al. (1995).

[18] These data are drawn from the Organization for Economic Cooperation and Development 2003, "Pollution Abatement and Control Expenditures in OECD Countries", Paris: OECD Secretariat.

oxides that may at present be approaching a peak in emissions. Conversely, particulate pollution peaked much earlier than the other pollutants, while lead has a dramatic drop in the mid-1970s. These raw U.S. data support the contention that environmental quality at first deteriorates and then improves with increases in income per capita.

Another interesting aspect of these figures is the breakdown of emissions by end-use category. Apart from some exceptions arising from the miscellaneous category the within-pollutant source of the emissions remains roughly constant in many of the figures. For example, consider SO2. Aggregate emissions follow an EKC pattern, but the components of fuel combustion and industrial processes do as well. A similar pattern is found in volatile organic compounds, but less so in the case of carbon monoxide which presumably is due to the change in automobile use over the period. In total the rough constancy in the within-pollutant sources of emissions suggests that the overall EKC pattern is not driven by strong compositional shifts.

Our finding of an EKC in the raw emission data is consistent with the recent flurry of formal empirical work linking per capita income and pollution levels. This empirical literature was fueled primarily by the work of Grossman and Krueger (1993, 1995) who found that, after controlling for other non-economic determinants of pollution, measures of some (but not all) pollution concentrations at first rose and then fell with increases in per capita income.[19] Their work is important in several respects: it brought the empirical study of aggregate pollution levels into the realm of economic analysis; it debunked the commonly held view that environmental quality must necessarily decrease with economic growth; and it provided highly suggestive evidence of a strong policy response to pollution at higher income levels.

Unfortunately, empirical research has progressed very little from this promising start. Subsequent empirical research has focused on either confirming or denying existence of similar relationships across different pollutants.[20] Subsequent research has shown that the inverse-U relationship does not hold for all pollution, and there are indications the relation may not be stable for any one type of pollution.[21] Since the empirical work on the EKC typically employs cross-country variation in income and pollution to identify parameters, it is perhaps not surprising that there are signs of parameter instability. This instability could arise from country specific differences in the mechanism driving the two processes, but very little, if any, work has gone into evaluating the various hypotheses offered for the EKC. This interpretation of the econometric problems is of course consistent with our finding that the raw US data offers a dramatic confirmation of Grossman and Krueger's cross-country results. Cross-country differences leading to parameter instability are of course irrelevant in a one-country context.

In its original application, the EKC was interpreted as reflecting the relative strength of scale versus technique effects. However, it is difficult to support this interpretation. To isolate either the scale or technique effect we need to hold constant the composition of output, but this is not typically done in this literature. Therefore, the shape of the EKC may reflect some mixture of scale, composition and technique effects.

Despite these criticisms, the major and lasting contribution of this literature is to suggest a strong environmental policy response to income growth. The EKC studies are generally supportive of the hypothesis that income gains created by ongoing growth lead to policy changes that in turn drive pollution downwards. However, as our discussion in later sections will show, an EKC is compatible with many different underlying mechanisms and is entirely compatible with pollution policy remaining unchanged in the face of ongoing growth.

While most studies do not present evidence that allows us to distinguish between the underlying mechanisms responsible for the EKC, two recent studies offer additional insights. Hilton and Levinson (1998) examine the link between lead emissions and income per capita using a panel of 48 countries over the twenty-year period 1972-1992. This study is important because it finds strong evidence of an inverted U-shaped relationship between lead emissions and per capita income, and then factors the changes in pollution into

[19] In addition to Grossman and Krueger (1993,1995), other early contributions are Shafik and Bandyopadhyay (1994), Selden and Song (1994), Hilton and Levinson (1998), Gale and Mendez (1996).

[20] See, for example, Selden and Song (1994), El-Ashry (1993), Harbaugh et al (forthcoming), Stern and Common (2001) and the surveys mentioned previously.

[21] Hilton and Levinson (1998) contains some of the most convincing evidence of an EKC. Harbaugh, et al. (forthcoming) examines the sensitivity of the original Grossman and Krueger finding to new data and alternative functional forms.

two different components. The first is a technique effect that produces an almost monotonic relationship between lead content per gallon of gasoline and income per capita. The second is a scale effect linking greater gasoline use to greater income.[22] This study is the first to provide direct evidence on two distinct processes (scale and technique effects) that together result in an EKC.

To interpret the empirical evidence as reflecting scale and technique effects one needs to rule out other possibilities. Although the authors do not couch their analysis in this context, their analysis implicitly presents the necessary evidence. First, they document a significant negative relationship between the lead content of gasoline and income per capita (post 1983). This relationship shows up quite strongly in just a simple cross-country scatter plot of lead content against income per capita. Since lead content is arguably pollution per unit output, it is difficult to attribute the negative relationship to much other than income driven policy differences.[23]

Second, the authors find a hump-shaped EKC using data from the post-1983 period, but in earlier periods they find a monotonically rising relationship between lead emissions and income. The declining portion of the EKC only appears in the data once the negative health effects of lead had become well known. The emergence of the declining portion in the income pollution relationship is very suggestive of a strong policy response to the new information about lead. The fact this only appears late in the sample makes it difficult to attribute the decline in lead to other factors that could be shifting the demand for pollution. For example if the declining portion of the EKC was due to increasing returns to scale in abatement, then it should appear in both the pre and post-1983 data and vary across countries being correlated with an appropriate measure of economic scale. If it was due to shifts in the composition of output arising naturally along the development path, why would it only appear in the post-1983 data? While it is possible to think of examples where these other factors are at play, the scope for mistaking a strong policy response for something else is drastically reduced in this study. The natural inference to draw is that the decline only occurs late in the sample because with greater information about lead's health effects, policy tightened and emissions fell.

A second important study is Gale and Mendez (1998). They re-examine one year of sulfur dioxide data drawn from Grossman and Krueger's (1993) study. The study does not offer a theory of pollution determination, but is original in investigating the role factor endowments may play in predicting cross-country differences in pollution levels. They regress pollution concentrations on factor endowment data from a cross-section of countries together with income-based measures designed to capture scale and technique effects. Their results suggest a strong link between capital abundance and pollution concentrations even after controlling for incomes per capita. Their purely cross-sectional analysis cannot, however, differentiate between location-specific attributes and scale effects. Nevertheless, their work is important because the strong link they find between factor endowments and pollution suggests a role for factor composition in determining pollution levels. That is, even after accounting for cross-country differences in income per capita, other national characteristics matter to pollution outcomes.

Combining our three stylized facts on pollution emissions presents us with an important question. How did aggregate emissions and emissions per unit output fall so dramatically in the U.S. without raising pollution abatement costs precipitously?

There are several possible explanations. One possibility is that ongoing changes in the composition of US output have led to a cleaner mix of production that has lowered both aggregate measures of costs and emissions. The downward trend in emissions per unit output shown in Figure 1 prior to the advent of the Clean Air Act suggests some role for composition effects. While changes in the composition of US output are surely part of the story, there are reasons to believe that they cannot be the most important part. Over the 1971-2001 period of active regulation by the EPA, total emissions of the 6 criteria air pollutants (Nitrogen Dioxide, Ozone, Sulfur Dioxide, Particulate matter, Carbon Monoxide, and Lead) decreased on average by

[22] Lurking in the background of this study is a composition effect operating through changes in the fleet of cars. This composition effect is not investigated in the paper, although it may be responsible for the jump in lead per gallon of gasoline use at low income levels shown in Figure 4 of the paper.

[23] To be precise we should note that since lead content per gallon is an average, and cars differ in their use of leaded versus unleaded gas, the composition of the car fleet is likely to be changing as well. Therefore, the fall in average lead content may reflect an income-induced change in the average age of the fleet (which would lower average lead content) plus a pure technique effect.

25%. Over this same period, gross domestic product rose 161% and pollution abatement costs have risen only slightly.[24] The magnitude of these emission reductions is too large for it to reflect composition changes alone.

To get a feel for the magnitudes involved note that if changes in the composition of output over the 1971-2001 period are to carry all the burden of adjustment, then we would set the changes in a_i to zero in (2). Then using the EPA's estimate of an average 25% reduction for E and the 161% increase for Y, we find that the weighted average of industry level changes must add up to - 186% change. This is just too large a realignment in the composition of industry to be credible.

It is also apparent from the figures that emissions for most pollutants have been falling since the early 1970s and as we saw earlier there are limits to how far aggregate emissions can fall via composition effects. Our earlier discussion of the static nature of the within-pollutant sources of emissions also argues against strong composition effects. Finally, there is little evidence that international trade is playing a major role in shifting dirty goods industries to other countries but stronger composition effects after the advent of federal policy in early 1970s would be necessary to explain the fall in emissions seen in the figures.[25] For this set of reasons it seems clear that composition effects alone cannot be responsible for the result.

Another possible explanation is that ongoing growth in incomes has generated a strong demand for environmental improvement. In this account, income gains over the post WWII period produce a change in policy in the early 1970s and usher in the EPA and the start of emission reductions. While this explanation fits with the decline in emission to output ratios and the lowered emissions since the early 1970s, it too cannot be the entire story. As we discuss in Section 4, if rising incomes are to be wholly responsible for the change in pollution policy, agents must being willing to make larger and larger sacrifices in consumption for improving environmental quality. For example, the theory models of Stokey (1998), Aghion and Howitt (1998), Smulders (2001), Lopez (1994), etc. all require a rapidly declining marginal utility of consumption to generate rising environmental quality and ongoing growth. But as Aghion and Howitt note:

> "Thus it appears that unlimited growth can indeed be sustained, but it is not guaranteed by the usual sorts of assumptions that are made in endogenous growth theory. The assumption that the elasticity of marginal utility of consumption be greater than unity seems particularly strong, in as much as it is known to imply odd behavior in the context of various macroeconomics models..." (p. 162.)

A rapidly declining marginal utility of consumption is required in earlier work because increasingly large investments in abatement are required to hold pollution in check.[26] This implies the share of pollution abatement costs in the value of output approaches one in the limit, which is inconsistent with available evidence.[27]

A final possibility is technological advance. Ongoing technological progress in production *and* abatement could simultaneous drive long run growth and hold pollution abatement costs in check. Technological progress in goods production could be the driving force for growth in final output, while technological progress in abatement allows emissions per unit of output to fall precipitously without raising environmental control costs skyward. In this explanation, income gains from ongoing growth are responsible for the onset of serious regulation in the 1970s, but the advent of regulation then brought forth improvements in abatement methods. As a consequence agents have not been required to make increasingly large sacrifices in consumption for improving environmental quality. As we show in section 4, this explanation is consistent with the predictions of both our Green Solow and Kindergarten models.

[24] These figures are from the EPA's Latest Findings on National Air Quality, 2001 Status and Trends, available at http://www.epa.gov/air/aqtrnd01/ published September 2002.

[25] See for example Antweiler et al. (2001) and Grossman and Krueger (1993).

[26] This restriction also implies a large income elasticity of marginal damage and many question whether the demand for a clean environment can be so income elastic. For example, McConnell (1997) argues that current empirical estimates from contingent valuation and hedonic studies do not support the very strong income effects needed.

[27] See the discussion in Aghion and Howitt (1998, page 160-161) and our discussion of abatement in Section 5 of Brock and Taylor (2003).

Before we proceed we should note that the stylized facts given thus far exclude a discussion of many other pollutants. By selecting only pollutants for which data is available we may have erred on the side of optimism since the measurement of pollutants is often a precursor to their active regulation. One important omission from the above is any discussion of air toxics such as benzene (in gasoline), perchloroethylene (used by dry cleaners), and methyl chloride (a common solvent). These are chemicals are believed to cause severe health effects such as cancer, damage to the immune system, etc. At present the EPA does not maintain an extensive national monitoring system for air toxics, and only limited information is available.[28]

Another omission is any discussion of the set of long-lived chemicals and chemical by-products that have found their way into waterways, soils and the air. These products have very long half-lives and produce serious health and environmental effects. Prominent among these in US history are DDT, PCBs, Lead, and most recently CFCs. Official estimates on emissions of these pollutants is difficult to find, but historical accounts and partial data indicate their emissions follow a pattern roughly similar to that of lead shown in Figure 8. As shown by the figure the history of lead is one of strong initial growth in emissions, followed by a rapid phase-out and virtual elimination. In fact, lead continues to be emitted in small amounts, whereas PCB emissions rose from very low production levels in the 1930s to millions of pounds per year of production in the 1970s, to end with a complete ban in 1979. Similarly, DDT was used extensively after WWII but banned in 1972. CFC production in the US rose quickly with the advent of refrigeration and air conditioning, but this set of chemicals now faces a detailed phase-out with CFC-11 and CFC-12 already facing a complete production ban. The salient feature of these accounts is strong early growth followed by quite rapid elimination.

A final failing of these data is that they are on emissions and not concentrations.[29] Concentration data is available for most of these data at least over the last 20 to 30 years, but the data is well known to be noisy and suffers from other problems related to comparability over time. Nevertheless most aggregate measures of air quality in cities have been improving over time. For example, data on the number of US residents living in counties that were designated non-attainment because of their failure to achieve federal air quality standards shows that over the 1986 to 1998 period these numbers have been falling quite dramatically for sulfur dioxide, nitrogen oxide, carbon monoxide, lead and PM10. The number of people living in counties who failed the federally mandated ozone air quality has however risen from 75 million in 1986 to 131 million in 1998.[30]

4 Some Illustrative Theory

4.1 The Green Solow Benchmark[31]

Every model relating economic growth to emissions or environmental quality has by construction made implicit assumptions regarding the strength of scale, composition and technique effects. These assumptions are often hidden in choices made over functional form, over the number of goods, the inclusion of finite resources, or in assumptions concerning abatement. Since we have data on the composition of output, its scale, and emissions per unit of output, it is often useful to divide models into categories according to their reliance on scale, technique and composition effects rather than model specifics like the number of goods, types of factors or assumptions over abatement. By dividing up the literature along these lines, we can weigh the relative merits of model's that rely exclusively on composition effects by looking at their strength in the data rather than by asking ourselves far less obvious questions such as are capital and resources good or poor substitutes or does abatement exhibit increasing returns.

The literature linking growth and pollution levels is immense starting with very early work in the 1970s by Forster (1973), Solow (1973), Stiglitz (1974), Brock (1977) and others, and culminating in the more recent

[28] See U.S. EPA, Toxic Air Pollutants, at www.epa.gov/airtrends/toxic.html.

[29] Note however that much of the empirical EKC work employs pollution concentration data as does Antweiler et al. (2001).

[30] For these data see U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, National Air Quality and Emissions Trend Report, 1998 (EPA, Research Triangle Park, NC 2000) and earlier trend reports.

[31] For a detailed exposition see Brock and Taylor (2004).

work investigating the Environmental Kuznet's curve such as Stokey (1998), Aghion and Howitt (1998), or Jones and Manuelli (2001). Although the earlier and late literatures differ greatly in their assumptions regarding the driving force behind growth, all models of economic growth must produce changes in scale, composition, or techniques that satisfy (1). Models that produce similar aggregate relationships between income and pollution often rely on different mechanisms to drive pollution downward. Because of these differences, they have other observable implications that we can use for evaluation.

To start our enquiry into the various mechanisms authors have employed to generate sustainable growth or an EKC prediction, we develop an augmented Solow model where exogenous technological progress in both goods production and abatement leads to continual growth with rising environmental quality. This is the simplest model to explore the importance of technological progress in driving down emissions per unit of output.[32]

Consider the standard one sector Solow model with a fixed savings rate s. Output is produced via a CRS and strictly concave production function taking effective labor and capital to produce output, Y. Capital accumulates via savings and depreciates at rate δ. We assume the rate of labor augmenting technological progress is given by g. All this implies:

$$\begin{aligned} Y &= F(K, BL) \\ \dot{K} &= sY - \delta K \\ \dot{L} &= nL \\ \dot{B} &= gB \end{aligned} \tag{13}$$

To model the impact of pollution we follow Copeland and Taylor (1994) by assuming every unit of economic activity, F, generates Ω units of pollution as a joint product of output. The amount of pollution released into the atmosphere may differ from the amount produced if there is abatement. We assume abatement is a CRS activity and write the amount of pollution abated as an increasing function of the total scale of economic activity, F, and the economy's efforts at abatement, F^A. If abatement at level A, removes the ΩA units of pollution from the total created, we have:

$$\begin{aligned} E &= \textit{pollution created} - \textit{pollution abated} \\ &= \Omega F - \Omega A\left(F, F^A\right) \\ &= \Omega F\left[1 - A\left(1, \frac{F^A}{F}\right)\right] \\ &= Fe(\theta) \text{ where } e(\theta) \equiv \Omega\left[1 - A(1, \theta)\right] \text{ and } \theta \equiv \frac{F^A}{F} \end{aligned} \tag{14}$$

where the third line follows from the linear homogeneity of A, and the fourth by the definition of θ as the fraction of economic activity dedicated to abatement.

The relationship in (14) requires several comments. The first is simply that (14) is a single output analog of (1) showing that emissions are determined by the scale of economic activity F, and the techniques of production as captured by $e(\theta)$. The second is that the production of output per se and not input use is the determinant of pollution. Since there is only one output, this means the composition effect must be zero. In a subsequent section we alter our formulation to consider pollution created by input use, but note in passing here that making pollution proportional to the employment of capital has no effect on our results. Finally, since F^A is included in F, even the activity of abatement pollutes.

To combine the assumptions on pollution in (14) with our Solow model, it is useful to assume the economy employs a fixed fraction of its inputs – both capital and effective labor – in abatement. This means the fraction of total output allocated to abatement θ is a fixed much like the familiar fixed saving rate

[32] A full description of the model together with supporting empirical work can be found in Brock and Taylor (2004), NBER working paper No. 10557.

assumption.[33] As a result, output available for consumption or investment Y, becomes $[1-\theta]F$. In addition we must adopt some assumption concerning natural regeneration. To do so we assume the quality of the environment evolves over time according to (6). Therefore, the evolution of the pollution stock is given by:

$$\dot{X} = Fe(\theta) - \eta X \tag{15}$$

Finally, since the Solow model assumes exogenous technological progress, we assume emissions per unit of output fall at the exogenous rate g^A. Putting these assumptions together and transforming our measures of output and capital into intensive units, our Green Solow model becomes:

$$\begin{aligned} y &= f(k)[1-\theta], \\ \dot{k} &= sf(k)[1-\theta] - [\delta + n + g]k \\ E &= BLf(k)e(\theta), \\ \dot{X} &= E - \eta X, \\ \dot{\Omega} &= -g^A \Omega \end{aligned} \tag{16}$$

where k is K/BL and y is Y/BL; i.e. capital and output measured in intensive units. The top line of (16) repeats the basic Solow model where net output is a fraction of total output. Taking θ as given, and assuming the Inada conditions, there is a k^* such that output, capital, and consumption per person all grow at rate g.

Using standard notation, direct calculation reveals that along the balanced growth path we must have $g_y = g_k = g_c = g > 0$. A potentially worsening environment however threatens this happy existence. From (16) it is easy to see that constant growth in environmental quality requires $g_x = g_E$. Since k^* is constant along the balanced growth path, the growth rate of emissions is simply:

$$g_E = g + n - g^A \tag{17}$$

which may be positive, negative, or zero. The first two terms in (17) represent the scale effect on pollution since output grows at rate $g + n$. The second term is a technique effect created not by greater abatement efforts, but because emissions per unit output fall via exogenous technological progress at rate g^A.

Therefore our requirements for sustainable growth are very simple in this model.

$$\begin{aligned} g &> 0 \\ g^A &> g + n \end{aligned} \tag{18}$$

Technological progress in abatement must exceed that in goods production because of population growth; and some technological progress in goods production is necessary to generate per capita income growth.

The Green Solow model, although very simple, demonstrates several important points. The first is that investments to improve the environment may cause only level and not growth effects. This is obviously true here since the growth rate of per capita magnitudes is explicitly linked to the rate of technological progress, but not to θ. By setting the time derivative of capital per effective worker to zero in (16) it is straightforward to show that a tighter environmental policy (higher θ) lowers output, capital and consumption per worker, but has no effect on their long run growth rates.

The implied income and consumption loss from a tighter policy is however quite small. Adopting a Cobb-Douglas formulation for final output with the share of capital equal to α shows that the ratio of consumption per person along the balanced growth paths of the economy adopting weak versus strong abatement is just:

$$\frac{c_w}{c_s} = \left[\frac{1-\theta_W}{1-\theta_s}\right]^{\frac{1}{1-\alpha}} \tag{19}$$

[33] We treat θ as endogenous when examining transition paths in section 5. It is possible that no abatement is optimal in some limited circumstances, but in models generating balanced growth this would imply every increasing pollution levels. In models without growth, such as Keeler et al. (1972), a Murky Age or Polluted Age equilibrium result is possible with θ set to zero.

since both economies grow at rate $g+n$. If weak abatement means adopting a share of pollution abatement costs in national output of .5%, and strong abatement means 10%, then consumption per person will differ by 16% along the balanced growth path (assuming capital's share in output is 0.35). Therefore, a twenty-fold difference in the intensity of abatement creates only a 16% difference in consumption per person!

The calculation however seems to imply that environmental policy cannot be much of a limit on economic growth. Recall though that for any given choice of abatement intensity, if we are to have ongoing growth, an improving environment, and a constant (relative) cost of environmental policy, then technological progress in abatement must be sufficiently strong. If technological progress is slower than that given in (18) then one of two things must happen. Either additional investments in abatement must be undertaken to maintain environmental quality, or environmental quality must decline. At this point however we should note that the strict concavity of A implies there are diminishing returns to greater and greater abatement efforts. From (14) we find that $e'(\theta) = -A_2 < 0$, but $e''(\theta) = -A_{22} > 0$. Therefore, in the absence of strong technological progress in abatement, growth in income per capita is only consistent with lower pollution levels if abatement grows over time.[34]

One final observation concerns the transition path of the model. Despite the fact that the intensity of abatement is fixed and there are no composition effects in our one good framework, simulations of the Green Solow model produce a path for income and environmental quality tracing out an Environmental Kuznets Curve. This surprising result is shown graphically in Figure 9 below. In Figure 9 we present the trajectories for two economies that are identical in all respects except for their abatement intensity. Each starts from an initially pristine environment and a small initial capital level. Strong abatement refers to a 10% share of output spent on abatement; weak abatement to a 0.5% share. The other parameters were chosen for the purposes of illustration. Per capita income grows at 1% along the balanced growth path, the population grows at 2% and the abatement technology improves at 5%. Note that these parameters ensure that growth is sustainable according to (18). Capital's share is set at 0.35, the savings rate is 10% and depreciation is 2%. Regeneration is set with $\eta = 0.1$ implying a 10% rate for any $X > 0$.

As shown, the environment at first worsens and then improves as the economy converges on its balanced growth path. Note that along the balanced growth path emissions fall and the environment improves at 2% per unit time, which is close to what the simulation delivers in its last periods.

This result follows for three reasons. First, the convergence properties of the Solow model imply that output growth is at first rapid but then slows as k approaches k^*. With a fixed intensity of abatement, pollution emissions grow quickly at first but slower later. Therefore, part of the dynamics is governed by the convergence properties of the neoclassical model.

Second, when we start at a pristine environment the effective rate of natural regeneration is zero. This is true because "nature" is at a biological equilibrium with $X = 0$. When production begins the environment deteriorates. At $X = 0$, the introduction of any emissions overwhelms the rate of regeneration and lowers environmental quality. As X rises, natural regeneration rises. This must be a feature of the regeneration function in order for $X = 0$ to be a stable biological equilibrium.

Third, we have assumed emissions fall along the balanced growth path.

Together the first two facts imply that at the outset of economic growth the rapid pace of growth swamps nature's slow or zero regeneration; but the economic growth rate slows and regeneration rebounds. As we approach the balanced growth path natural regeneration must overwhelm the now less rapid inflows of pollution. The environment improves.

It is important to recognize that this result is more general than our Cobb-Douglas technology and instead relies on quite general properties of production and growth functions. To verify write the dynamic system governing k and X as:

$$\begin{aligned} \dot{k} &= sf(k)[1-\theta] - [\delta + n + g]k \\ \dot{X} &= c_o \exp[-c_1 t] f(k) - \eta X \end{aligned} \tag{20}$$

[34] A rising intensity of abatement will lower growth rates introducing other problems in meeting the sustainability criteria. We leave this issue for the next section.

where $c_0 > 0$ and $c_1 = g^A - (g+n) > 0$ are positive constants. To show the environment must at first worsen evaluate the accumulation equation for X at $t = 0$. Since the environment is initially pristine, $X(0) = 0$, and the initial capital stock cannot be zero so $k(0) > 0$. Substituting these values into (20) and evaluating shows the environment must at first worsen. X has to be growing at least initially. To examine the rest of the transition path recall that $k(t)$ is increasing in time until it reaches k^* because this is, after all, a Solow model. Using this fact, we can bound the path for X by noting:

$$\dot{X} = c_o \exp[-c_1 t] f(k) - \eta X < c_o \exp[-c_1 t] f(k^*) - \eta X \tag{21}$$

Therefore for any $t > 0$, $X(t)$ must be below the solution to the ordinary differential equation: $\dot{X} = c_o \exp[c_1 t] f(k^*) - \eta X$, $X(0) = 0$. This ordinary differential equation has a closed form solution showing $X(t)$ tends to zero as t goes to infinity. Using the inequality in (21) we can conclude that pollution must fall along its trajectory.[35]

This explanation for the EKC is entirely distinct from those offered in the literature. There are no composition effects, no increasing intensity of pollution abatement, no increasing returns to abatement, no evolution of the political process, and no international trade. The result follows primarily from the mechanics of convergence coupled with the dynamics predicted by a standard natural regeneration function. Moreover from (14) it is easy to see that emissions per unit of GDP falls *both* during the transition and along the balanced growth path at the constant rate g_A (recall Figure 1 at this point). This is quite surprising because both output and emissions growth varies over time, with the level of emissions tracing out an EKC. Since θ is fixed throughout, the share of pollution abatement costs in value-added is constant. Therefore, although very simple, the Green Solow model matches three important features of the data: declining emissions to GDP ratios, the EKC, and pollution abatement costs that are roughly constant over time.

The Green Solow Model bears a family resemblance to several papers examining the growth and pollution link within a neoclassical framework. Forster (1973) examines a neoclassical model with zero population growth and no technological progress in either abatement or production. His main result is that steady state consumption per person and capital per person are lower when society invests in pollution control since these controls lower the net return to capital.

Although Forster's assumptions on abatement and pollution creation are different from ours, we can reproduce his main results in our fixed savings rate setting by adopting his assumptions of $g_A = g = n = 0$. When we do so we find steady state capital per person and consumption per person both fall with increases in θ while pollution is lowered. Forster's work is important because it was perhaps the first examination of optimal pollution control in a neoclassical setting.

The Green Solow model is also similar to the neoclassical model adopted in Stokey (1998), but differs in that Stokey gives no role to technological progress in abatement. As a result increasing abatement intensity must carry the day in reducing pollution. Stokey also generates the EKC prediction but her result follows from a change in pollution policy along the transition path. Her simulations of the model must however to some extent reflect the same dynamic forces we have identified since the model is neoclassical and the evolution equation for the environment is identical.

More closely related work is Bovenberg and Smulders (1995). In their endogenous growth formulation "pollution augmenting technological progress" holds pollution in check *and* drives long run growth. In their two-sector model, ongoing investments in the knowledge sector raise the productivity of pollution leading to a balanced growth path with a constant level of environmental quality. Again our Green Solow model reproduces the flavor of their results. Setting $n = 0$ to mimic their zero population growth assumption, and assuming $g = g_A$ to mimic the identical rates of technological progress found in both sectors, we find from (22) that emissions are constant along the balanced growth path and output per person grows at rate g.

This similarity should not be all that surprising because Bovenberg and Smulders' "pollution augmenting technological progress" is very similar to our technological progress in abatement. To see why divide both sides of our emissions function in (14) by Ω, and then employ the monotonicity of A in θ to invert the

[35] With a further assumption on technology we can ensure the EKC must be single humped. Our Cobb-Douglas formulation adopted in the figure is covered by our assumption.

intensive abatement function and solve for $[1-\theta]$. Use this to write net output $Y = F[1-\theta]$ available for consumption or investment as:

$$Y = G\left(F\left(K, BL\right), AE\right) \tag{22}$$

where $A = 1/\Omega$, and G is both CRS and concave. Hence "pollution augmenting technological progress" is equivalent to our technological progress in abatement.

4.2 Intensifying Abatement: The Stokey Alternative

We now amend our Green Solow model to incorporate a role for intensified abatement to lower pollution levels. In the model above reductions in pollution came about solely because of changes in the emission technology and not because society allocated a greater share of its resources to pollution prevention. In an important paper Nancy Stokey [Stokey (1998)] presented a series of simple growth and pollution models to investigate the links between the limits to growth and industrial pollution. She examined the ability of these models to reproduce the results of empirical work finding an Environmental Kuznets Curve and investigated how an active environmental policy may place limits on growth. An important feature of Stokey's analysis was its dependence on increased abatement and tightening regulations to drive pollution downward.

Her analysis contains two contributions. The first is a simple explanation for the empirical finding of an Environmental Kuznets curve. Like Lopez (1994) before her, and Copeland and Taylor (2003) after, Stokey shows how an income elastic demand for environmental protection can usher in tighter regulations and eventually falling pollution levels. This assumption on tastes, together with certain assumptions on abatement, succeeds in generating a first worsening and then improving environment as growth proceeds.

Stokey's second contribution was to investigate whether there are limits to growth imposed by regulating industrial pollution. In section 5 we discuss her analysis within an AK framework; here we focus on her work within the neoclassical model that formed the bulk of her paper. To do so we make the smallest departures possible from the Green Solow model. We again take the savings rate as fixed, but allow the intensity of abatement to vary over time. Since we are primarily interested in feasibility rather than optimality, our fixed savings rate assumption will simplify the analysis at little or no cost. Stokey assumed zero population growth, exogenous technological progress in goods production, a Cobb-Douglas aggregator over capital and labor in final goods production, and adopted an abatement function drawn from Copeland and Taylor (1994). In Stokey's analysis an optimizing representative agent determine savings and abatement decisions.

Although it is not obvious from Stokey (1998), a process of pollution abatement is implicit in her analysis. In Stokey's formulation the planner chooses a consumption path and the techniques of production as indexed by "z". The choice of techniques determines the link between potential output, F, and final output Y available for consumption or investment. The two are related by $Y = Fz$; while aggregate emissions are given by $E = Fz^{\beta}$ for some $\beta > 1$. To see how this choice of "techniques" is really one over abatement intensity make the change of variables $(1-\theta) = z$, and then let $\mathrm{e}(\theta) = (1-\theta)^{\beta}$ for $\beta > 1$. It is now easy to see that the "techniques" chosen by the planner correspond to choices over the abatement intensity θ. The resulting $e(\theta)$ is just a specific form of an emissions function coming from the assumptions of constant returns to abatement and pollution being a joint product of output. Since θ is in principle observable, we conduct our analysis in this unit.

Our amended model assumes zero technological progress in abatement, and to follow Stokey adopts the specific emissions function given above and a Cobb-Douglas aggregator over factors. The model is described by:

$$\begin{aligned} Y &= F[1-\theta] \\ F &\equiv K^{\alpha}(BL)^{1-\alpha} \\ \dot{K} &= sY - \delta K \\ \dot{L} &= nL \\ \dot{B} &= gB \\ E &= Fe(\theta) \equiv F[1-\theta]^{\beta} \end{aligned} \tag{23}$$

To examine the feasibility of balanced growth with a non-deteriorating environment we start with the last equation in (23) giving emissions and log differentiate to find:

$$G_E = \alpha G_k + (1-\alpha)(g+n) + G_{e(\theta)} \tag{24}$$

recall G_E must be zero or negative or else the environment deteriorates. In the absence of technological progress in the emissions function, this implies the growth rate of emissions per unit of output $G_{e(\theta)}$ must be negative. To identify what this may imply for growth, we eliminate G_K. To do so, note balanced growth requires Y/K constant. Divide both sides of the final goods production function by Y^{α}. Rearrange and log differentiate with respect to time to find:

$$G_Y = G_K = g + n + \frac{G_{e(\theta)}}{(1-\alpha)\beta} \tag{25}$$

where G_Y is the growth rate of final output. At this point it is worthwhile to note that final output growth is reduced by active abatement since $G_{e(\theta)}$ must be negative.

To determine the evolution of emissions along the balanced growth path substitute (25) into (24) and rearrange slightly to obtain:

$$G_E = g + n + \frac{G_{e(\theta)}}{(1-\alpha)\beta} + \frac{(\beta-1)G_{e(\theta)}}{\beta} \tag{26}$$

The first two terms of this expression, $(n+g)$, represent the scale effect of growth. They represent the growth rate of emissions that would arise along the balanced growth path if θ was held constant. This is clear from (23) since if θ is constant $G_{e(\theta)} = 0$.

The final two terms in (26) represent the technique effect created by lowering emissions per unit output along the balanced growth path. This technique effect is itself composed of two parts. The first component is the reduction in the growth rate of final output caused by the diversion of resources to abatement. Since θ is increasing along the balanced growth path, the growth rate of F exceeds that of final output by this amount.[36] Therefore, this component of the technique effect lowers pollution by slowing down the growth rate of final output (recall (25)).

The second component of the technique effect is the reduction in emissions per unit of final output created by abating more intensively. This is the standard component identified in static models. This component of the technique effect need not be as large as previously to lower pollution. To see this solve (26) for the rate at which emissions per unit of output must fall to drive emissions downward. Algebra yields:

$$-G_{e(\theta)} > \left[\frac{g+n}{1+[\alpha/[(1-\alpha)\beta]]}\right] \tag{27}$$

which is smaller than the minimum rate of $(g+n)$ needed in (17). Not surprisingly because abatement has a negative effect on growth rates, it has less of a role to play in lowering emissions per unit of output. Therefore in set-ups where abatement is responsible for pollution reduction, the burden is shared across two margins: abatement lowers growth rates and abatement also lowers emissions per unit output.

These two roles for abatement now introduce the possibility that a sustainable growth path may not exist. To see why, note that the reduction in growth created by an ever tightening environmental policy is very similar to the growth drag found in models with either fixed land or exhaustible natural resources.[37] In the case with fixed or declining resources the ratio of resource use to effective labor falls along the balanced growth path and this lowers growth rates. The same is true here once we make the right translations. To see this parallel, use the final goods production function and the emissions function to write net output as if pollution were an input into production. Doing so we find:

$$Y = F^{1-\frac{1}{\beta}} E^{\frac{1}{\beta}} \tag{28}$$

[36] The growth rate of θ and $e(\theta)$ are related by $G_{e(\theta)} = -\beta[\theta/(1-\theta)]G_{\theta}$.

[37] An excellent review of growth drag is contained in Jones (2002; chapter 9).

Along the balanced growth path E must fall while F grows; therefore the reduction in E works very much like the exhaustion of a resource that lowers growth. It is now apparent that while (27) tells us that the decline in emission intensity must be sufficiently fast to lower emissions; equation (25) tells us this same magnitude cannot be too large if we are to have positive growth in income per capita. Solving (25) for the implied restriction and combining with (27) yields, after some manipulation:

$$g > -G_{e(\theta)} > \frac{g+n}{\alpha + (1-\alpha)\,\beta} \tag{29}$$

The range given by this inequality defines the set of emission intensity reductions that are consistent with declining emissions and positive per capita output growth: i.e. sustainable growth. If we recall that $\beta > 1$, then it is straightforward to see that the feasible region is not empty when there is zero population growth. When $n > 0$ the region may not exist. By equating the two sides of (29) we can solve for the relationship between population growth and parameters that must be true for a sustainable growth path to exist. Algebra gives us:

$$g\left[(1-\alpha)\,(1-1/\beta)\right] > n\left[\frac{1}{\beta}\right] \tag{30}$$

The left hand side of (30) is exactly labor's share in final goods production (use (28)and the definition of F) times the rate of labor augmenting technological progress, g. The right hand side is exactly emissions share in final output, $1/\beta$, times the rate of population growth, n. The intuition for this condition is straightforward, and is identical to that we give later in a model where exhaustible energy resources create drag.

The left hand side of the expression represents the Solow forces of technological progress raising growth to the extent determined by labor's share and the rate of progress. The right hand side could be called the Malthusian forces since they capture the impact of diminishing returns caused by a falling ratio of emissions to effective labor along the balanced growth path. These forces are stronger the more important are emissions in the production of final output, and stronger the faster is population growth.

If the inequality in (30) goes the other way then we have two choices. Either per capita income growth is negative, or per capita income growth is positive but emissions rise. In either case we do not have sustainable growth according to our definition.

This observation of course suggests we follow the path of earlier authors and calculate the growth drag due to pollution policy. For example, Nordhaus (1992) adopts a model similar to (28) with emissions E replaced by either land or an exhaustible natural resource and then generates estimates for the drag caused by finite land and natural resources. But without a formal framework in which to estimate the long run growth impact of tighter environmental policy, Nordhaus resorts to estimates of contemporaneous expenditures on abatement to calculate future costs of pollution control.

We can go further here, although our methods are far from ideal. To generate an estimate for the growth drag caused by environmental policy we need estimates of β, α and $G_{e(\theta)}$. We note using (23) that $G_{e(\theta)} = G_{E/Y}\beta/[\beta-1]$ where $G_{E/Y}$ is the observable growth rate of emissions per unit of final output. For various measures of E it is shown in Figure 1. We take capital's share of production, α to be 0.35. To eliminate the parameter β write emissions per unit of output, using (23) as $E/Y = (1-\theta)^{\beta-1}$. Since we have data on emissions, final output and pollution abatement costs we could in theory estimate β. Using this estimate we could then calculate the growth drag due to pollution policy. Since our purpose is not to provide definitive answers but rather suggest a methodology, take the log of E/Y and differentiate with respect to time to find:

$$\beta = 1 - \frac{G_{E/Y}}{G_\theta}\left[\frac{1-\theta}{\theta}\right] \tag{31}$$

where G_θ is the growth rate of the pollution abatement cost share, and θ the average pollution abatement cost share over the period in consideration. Now use (31) to eliminate β and rewrite (25) as:

$$G_Y = g + n - \frac{G_\theta}{(1-\alpha)}\frac{\theta}{(1-\theta)} \tag{32}$$

PAC share percentage increase per year	2.5	5.0	7.5
Pollution abatement costs share 1970	1.0	1.0	1.0
Pollution abatement cost share 2000	2.1	4.3	8.8
Average $\theta/[1-\theta]$ across period	1.57	2.72	5.15
Growth drag percentage	0.06	0.2	0.5

Table 1: The drag pollution policy on growth (percentages)

The drag due to environmental policy is now directly linked to observable measures: the share of pollution abatement costs in the value of overall economic activity, and the percentage growth rate of this measure. To investigate what a reasonable magnitude of growth drag maybe, we report in the table below a series of illustrative calculations. Recall that the share of pollution abatement costs in either manufacturing value-added or GDP is small – on the order of 1 or 2%. In certain industries it can of course be much higher. Take 1970 as the base year and set the pollution abatement costs share in that year at 1%. Then applying growth rates of 2.5 to 7.5% per year in this cost share, we obtain with the help of (32), the following results.

The first column assumes the share of pollution abatement costs in the value of output rises from 1% to a little over 2% in thirty years. The other columns report larger increases for illustrative purposes, although they are far in excess of the historic increases as shown by our data in Figure 2. A striking feature of the table is that the drag due to environmental policy is very small except in extreme cases. When pollution abatement costs rise from 1% to a little over 2% in 30 years, the drag on growth is only 6 hundredths of 1% point. When pollution abatement costs grow by 5% per year, the policy reduces growth by 0.2%. If costs grow by the extremely large 7.5% per year, drag is now $\frac{1}{2}$ of 1% point which is significant. Note that growth in per capita income, $G_Y - n$, over the last 50 years is approximately 2% per year; therefore the last column would predict an ever strengthening environmental policy that raises the share of pollution abatement costs in value-added by 7.5% year would reduce per capita income growth by 25%.

To a certain extent the relatively small effects in Table 1 are not that surprising. If pollution abatement costs as a fraction of value-added are in the order of 1%, it is difficult to see how even relatively large percentage increases in their level would lower growth tremendously. To go slightly further, note from (31) that if $G_{E/Y}$ and G_θ are of the same magnitude, then it is easy to see that β is approximately $1/\theta$.[38] This implies the share of emissions in final production in (28), $1/\beta$ is on the order of 0.01 or 0.02.And if pollution emissions are such an unimportant input into the production of final output, then drag from any reduction in emissions over time must also be small.

Despite the optimistic results in Table 1 concerning growth rates, models that rely on active abatement often contain the prediction that abatement becomes a larger and larger component of economic activity. This is a direct consequence of two facts. The first is that for emissions to fall, emissions per unit of output must shrink continuously with ongoing growth. The second is that with constant returns to abatement, lowering emissions per unit output comes at increasing cost. As a consequence, an implication of an exclusive reliance on abatement is that abatement costs rise along the growth path to eventually take up most of national product. To verify this, return to our simple example and note $G_{e(\theta)}$ is constant along the balanced growth path. Using our specific emission function in (23), we know:

$$G_{e(\theta)} = -\beta \left[\frac{\dot{\theta}}{1-\theta} \right] \tag{33}$$

Solving this differential equation for θ yields:

$$\theta(t) = 1 - (1 - (1 - \theta(0)))\, e^{\left(G_{e(\theta)}/\beta\right)t} \tag{34}$$

[38] This may not be such a bad assumption. In Figure 1 it appears that the growth rate of emissions per unit output for each pollutant may be roughly constant over the last 50 years. The important point is that the two growth rates are of a similar magnitude and not necessarily equal.

starting from some $\theta(0)$ near the balanced growth path we see that as time goes to infinity θ goes to one because $G_{e(\theta)} < 0$. Abatement must take up a larger and larger share of national product as time progresses. This is an uncomfortable conclusion in light of the data we have already presented showing a relatively weak increase in abatement over time. In addition the reader may wonder why it is that agents would willingly make such sacrifices in final consumption necessary for such a large abatement program.

At this point it is useful to refer to Stokey (1998) explicitly for an answer since Stokey's analysis shows that consumer's are indeed willing to make the sacrifice needed in net output to lower pollution albeit under certain conditions. Specifically, by adopting a CRRA utility function Stokey shows emissions fall along the balanced growth path if and only if the elasticity of marginal utility with respect to consumption exceeds one. Only if consumers valuation of consumption falls quickly are they willing to take a smaller and smaller slice of (an ever expanding) national income as growth proceeds.

Stokey's analysis also allows for a more theoretically based growth drag calculation. By adopting specific functional forms, Stokey solves for the growth rate of final output and emissions in terms of primitives. By rearranging slightly and recasting these results in terms of our notation we find the growth rate of output per person and overall emissions are just:

$$\begin{aligned} G_y &= g - g\left[\frac{\sigma+\gamma-1}{\sigma+\gamma-1+(1-\alpha)(\beta-1)\gamma}\right] \\ G_E &= \frac{1-\sigma}{\gamma}G_y \end{aligned} \tag{35}$$

where σ is the elasticity of marginal utility in the CRRA utility function, $\gamma \geq 1$ is a measure of the convexity of damages from pollution, α is capital's share, g is the exogenous rate of labor augmenting technological progress. There is zero population growth so $n = 0$.[39]

In comparison to our simple example the drag of environmental policy is now directly linked to the primitives of tastes and technology although it reflects similar forces at work. For example, by rearranging we can isolate the percentage reduction in growth caused by active pollution policy:

$$G_y = g\left[1 - \frac{1-G_{E/Y}}{1-G_{E/Y}+(1-\alpha)(\beta-1)}\right] \tag{36}$$

the greater is the rate of reduction in emissions per unit output, GE/Y, and the larger is emissions share in final output $1/\beta$, the greater will be the drag. This is the same set of forces we found using our simpler framework.

We can of course estimate growth drag in this optimizing framework as well. In order to replicate the Environmental Kuznet's curve Stokey adopts a set of parameters for all the primitives we need. In doing so, the model predicts the EKC found in empirical work, but using these same values for capital's share, the abatement technology, etc. we find that growth drag is an unbelievable 60% of potential growth. Using the parameter specification chosen to mimic the EKC, growth in income per capita in the absence of pollution policy is 4% per year.[40] But using (35) growth is actually approximately 1.6% per year with active pollution policy; therefore, growth in income per person is slowed by 60% from what it would be in the absence of environmental concerns. This is clearly far too high.

If we lower the elasticity of marginal utility to approach the lowest limit consistent with falling pollution (σ approaching 1), and set $\gamma = 1$, then drag hits its minimum. But even in this case, drag is almost 55% of potential growth. The problem with these calculations is our assumption of $\beta = 3$, which implies a share of pollution emissions in final output of 1/3 which is clearly far too high. Altering β to values similar to those used in our growth drag calculations suggests a much smaller drag.

For example, from Figure 1 it is apparent that 3 of the US criteria pollutants had an emissions per unit of output in 1998 that were just 1/10 of their value in 1940. This implies a growth rate of approximately

[39] This is found by rearranging (3) page 14 of Stokey (1998). To rewrite the equation in our set up we need to note the rate of labor augmenting technological progress would be $g/(1-\alpha)$ which we write as g in the above.

[40] This is just the effective rate of labor augmenting technical change which is $g/(1-\alpha) = 0.024/0.6$ in Stokey's notation.

-4% per year from these pollutants. Assuming the share of emissions in final output is 0.02, β is 50, and with a capital share of 0.35 we find the percentage reduction predicted by (36) to be just 0.03. Therefore a 2% growth rate would be reduced to just 1.94% because of the drag of environmental policy.[41]

We would hasten to add however that these calculations are purely for illustration. They demonstrate how the growth drag due to environmental policy may be calculated from primitives on technologies, abatement costs, knowledge of historic growth, and emission levels. We leave it to future research to develop and refine these methods to generate estimates of the growth drag due to environmental policy.[42]

Many other papers rely on an active role for abatement in lowering pollution levels, and therefore must contain predictions for both the drag of environmental policy and the evolution of pollution abatement costs over time. In some work, abatement is specified differently so that it escapes diminishing returns by assumption. For example, early work by Keeler et al. (1972) examines no growth steady states and assumes foregone output is the *only input* into abatement. As a result of this assumption, marginal abatement costs are constant in their formulation. Even with constant marginal abatement costs they find that when abatement is not very productive a "Murky Age" equilibrium arises: abatement is not undertaken and emissions are high in the steady state. Alternatively, when abatement is very productive in reducing emissions, the steady state is given by a Golden Age equilibrium with active abatement and lower emissions.

Other related work appears in Lopez (1994) and Copeland and Taylor (2003). In these contributions an optimizing social planner chooses the optimal level of abatement but factor supplies and technology are taken as parametric in their exclusively static analyses. Both adopt formulations where abatement is a constant returns activity using conventional inputs and examine how once for all growth in either technology or factor endowments affect pollution levels.[43] When growth is fueled by neutral technological progress, Copeland and Taylor show that emissions fall with this source of growth if the elasticity of marginal damage from pollution exceeds one. In a CRRA framework this corresponds to the condition Stokey derived of $\sigma > 1$.

In contrast when growth occurs by primary factor accumulation alone, then Lopez (1994) shows that whether pollution rises or falls now depends on both the elasticity of substitution between factor inputs and the income elasticity of marginal damage. If the elasticity of substitution between primary factors and emissions is large, then emissions fall quite easily. When production is Cobb-Douglas, Lopez's condition is identical to that of Stokey and $\sigma > 1$ generates the result that emissions fall with growth.

Together these contributions demonstrate that an improving environment and rising incomes are surely feasible in a standard neoclassical model where abatement is a constant returns activity. This path is also optimal under certain conditions. But by relying exclusively on changes in the intensity of abatement to lower pollution levels consumers must be willing to make rather large sacrifices for a cleaner environment over time. It is in fact this rather large willingness to sacrifice for a cleaner environment that leads to regulation in the first place.

In Stokey (1998), regulation is at first not present as the shadow value of capital is too high and the shadow value of pollution too low when growth begins for the planner to allocate any output to abatement. An important input into this decision is that the marginal product of the first unit of abatement is bounded above even at zero abatement.[44] As a result, no abatement is undertaken $\theta = 0$ and pollution rises lockstep with output. Once the environment has deteriorated sufficiently and the now larger capital stock has depressed its shadow value, active abatement begins. There is then a transition period and the economy approaches the balanced growth path described previously.

[41] The problem with this set of parameters is that the output elasticity of emissions in production is far too high at 1/3. If the regulator used pollution taxes to implement the social optimum, this implies that at all periods of time the share of pollution taxes in value-added would be 1/3. Setting β much higher generates numbers closer to those we reported in Table 1.

[42] Other methods used to estimate the impact of tighter pollution policy on growth include the use of quite detailed computable general equilibrium models of the U.S. economy, econometric studies, and more aggregative data exercises like the one we just conducted. The results from these studies are quite different. For example, Jorgenson and Wilcoxen (1990) build a 35 industry model of the U.S. economy to estimate the impact of pollution abatement costs and motor vehicle emission standards on overall output and growth. They find that output growth in the U.S. was reduced by almost 0.2% over the 1973-1985 period by these environmental policies, and in level terms U.S. real GDP is lower by a quite significant 2.6%.

[43] Lopez (1994) does not present an abatement function per se but it is implicit in his use of the revenue function listing primary factors and emissions as productive factors.

[44] We will show this in section 5.

This explanation for the EKC is quite persuasive. It links rising income levels with a lower shadow cost of abatement and a higher opportunity cost of doing nothing. It captures the idea that policy responds positively to real income growth and generates an EKC is a straightforward way. We have already seen however that a further implication of the model is an ever-rising pollution abatement cost share that may be inconsistent with the data. In addition we should note that this explanation predicts a constant emissions to output ratio prior to the regulation phase when emissions are rising. After regulation begins, the emissions to output ratio falls and does so at a constant rate along the balanced growth path (see (35)). Figure 1 however shows the emission to output ratio was falling long before emissions peaked in Figures 3 through 8. Therefore, using this data as our guide the model misses the long reduction in emissions per unit output that occurred in the US prior to peak pollution levels being achieved.

These observations suggest, at the very least, that other forces are simultaneously at work and partially responsible for the falling emissions to output ratio and roughly constant control costs in the U.S. historical record. One natural candidate is of course changes in the composition of output towards less energy intensive, and hence less pollution intensive, goods. Much has been made recently about the dematerialization of production and its environmental consequences, and hence we now turn to examine a model relying on just these effects.

4.3 Composition shifts: The Source and Sink Model

There are several ways to escape a worsening environment as economic growth proceeds. One possibility is for technological progress in abatement to lower pollution levels as shown in the Green Solow model; another is intensified abatement as shown by the Stokey Alternative. A third method is to alter the composition of output or inputs towards less pollution intensive activities. In this section we investigate the implication of changing energy use in production. Much of current concern over pollution arises from energy use and hence if the economy as a whole could conserve on energy this would have important implications for environmental quality. But raising energy efficiency per unit of output comes at some cost because energy is a valuable input and constraining its use will lower overall productivity. These losses must be compensated for by increases in capital, effective labor or new technology if growth is not to be slowed. Therefore, solving our pollution problems by altering an economy's input mix may introduce significant drag. These growth concerns are of course one of the major reasons why many countries have delayed ratification of the Kyoto protocol; and why many developing countries refuse to sign the agreement.

While many models investigating the growth and pollution relationship rely on compositional changes to lower pollution levels, few make the role of energy explicit in their analyses. For example, Copeland and Taylor (2003) present a "Sources of Growth" explanation for the Environmental Kuznets curve arguing that if the development process relies heavily on capital accumulation in the earliest stages and human capital formation in later stages, these changes will alter the pollution intensity of production so that the environment should at first worsen and then improve over time. Related empirical work in Antweiler et al. (2001) finds growth fueled by capital accumulation is necessarily pollution increasing, while growth fueled by neutral technological progress lowers pollution levels. Behind these results is presumably a link between the different types of growth, energy use, and emissions.

Similarly, in Aghion and Howitt (1998)'s analysis of long run growth and environmental outcomes, their clean capital - knowledge - takes on a larger and larger role in growth in the long run and this too creates an eventually improving economy. But since they adopt the same assumptions on abatement as Stokey (1998), even with a changing composition of output large increases in abatement must made to hold pollution down to acceptable levels.

In most of these formulations the link to energy use is at best implicit with the reader having to interpret capital or other productive factors in a broad way to include energy or other natural resources. One of the major accomplishments of the early resource literature was to identify how and when finite resources impinge on the growth process. By ignoring the role of exhaustible resources in generating pollution, we run the risk of making pollution reductions look relatively painless because these analyses will miss the induced drag on economic growth created by lower energy use. In this section we make the connection between energy use,

growth and environmental outcomes precise by combining earlier models of growth and exhaustible resources with newer models examining the pollution and growth link. By doing so we demonstrate how some of the results of the earlier 1960s and 1970s literature on natural resources and growth have relevance today.

One of the major research questions of the earlier 'limits to growth" literature was the extent to which exhaustible natural resources impinged on growth. Seminal contributions by Solow (1974) and Stiglitz (1974) showed that growth with exhaustible resources was indeed possible, although it required a joint restriction on the rate of population growth, technological progress and the share of natural resources in output. There are two well-known results from this literature.

The first, due to Solow (1974), is that a program of constant consumption is feasible even with limited exhaustible resources and a constant population if the share of capital in output exceeds the share of resources in final output. This observation led to a consideration of the optimal rate of savings to maximize the constant consumption profile. The answer was provided by John Hartwick (1977) and embodied in the now-famous Hartwick's rule: invest all the rents from the exhaustible resource in capital and future generations will be as well off as the currently living despite the asymptotic elimination of natural resources.[45]

The second result, due to Stiglitz (1974), is that growth in per capita consumption is possible with positive population growth if the rate of resource augmenting technological progress exceeds the population growth rate. Our formulation will also yield a similar restriction on technological progress to generate positive per capita output growth, but in addition we add the further restriction that environmental quality improves. Therefore, even when growth with exhaustible resources is feasible in terms of generating positive output growth (as required by Stiglitz (1974)), it may be unsustainable because this same plan implies rising pollution levels.

We remain as close as possible to our earlier formulation while introducing a role for natural resources. We make two important changes. First, we introduce energy as an intermediate good. The intermediate good "energy" is produced from an exhaustible natural resource, R, capital, and labor via a CRS and strictly concave production technology. Final output (used for investment or consumption) is then produced via capital, labor and the energy intermediate. To keep things simple we assume both production functions are Cobb-Douglas, and to remain consistent with our earlier formulations we assume technological progress is labor augmenting.[46]

Our second change is to assume pollution is produced via energy use, and not the overall scale of final goods production. In doing we sever the strong link we had thus far between pollution and final output by making pollution the product of input use. We retain our earlier assumption that pollution can be abated, but take the fraction of resources devoted to abatement as constant. We have already shown in the Stokey Alternative that increasing abatement creates drag on economic growth; here we show that even with the abatement intensity fixed, a move towards less energy intensive production lowers growth while it reduces the growth rate of emissions.

With these assumptions in hand, the production side of the economy becomes:

$$\begin{aligned} Y &= K_y^{b_1} \left(BL_y\right)^{n_1} I^{b_2} \\ I &= K_I^{b_3} \left(BL_1\right)^{n_2} R^{b_4} \left[1-\theta\right] \end{aligned} \tag{37}$$

where I is the energy intermediate, θ is the fraction of the energy industry's activities devoted to abating pollution, R denotes the flow of resources used per unit time, and subscripts denote quantities of capital and labor used in final good production, Y, and the intermediate good energy, I.

Capital and labor has to be allocated efficiently across the two activities – intermediate and final good production. It is straightforward to show that this implies a constant fraction of the capital stock is employed in intermediate good production and the remainder in final goods. The same is true for labor. This allows

[45] Adopting Hartwick's rule in our source-and-sink model leads to increasing utility over time as the environment improves with resource exhaustion. Proof available upon request.

[46] This implicitly assumes that energy is not an essential input into production; i.e. energy per unit of output is not bounded below.

us to aggregate and rewrite the production function for final good output as follows:

$$Y = K^{a_1} (LB)^{a_2} R^{a_3} [1-\theta]^{b_2} \tag{38}$$

which is necessarily CRS with $a_1 + a_2 + a_3 = 1$.[47]

To complete the model we add equations governing the labor force and technology growth, the relationship between extraction and the resource stock, S, plus our abatement assumptions linking emissions to energy use, I. These conditions are:

$$\begin{aligned} \dot{K} &= sY - \delta K \\ \dot{L} &= nL \\ \dot{B} &= gB \\ E &= I\Omega e(\theta) = I\Omega [1-\theta]^{\beta} \\ \dot{S} &= -R \end{aligned} \tag{39}$$

where $e(\theta)$ measures the flow of emissions released per unit of energy used. It is instructive at this point to rewrite the emissions equation to focus on the role of a changing composition of inputs in determining pollution levels. To do so define the variable $\chi = I/Y$ which is the ratio of energy use to final good production, or what is commonly called the energy intensity of GDP. Then rewriting the emissions function we find:

$$E = \chi Y \Omega e(\theta) \tag{40}$$

A change in emissions can now come from any one of three sources: scale effects via changes in final good output Y; composition effects coming from changes in the energy intensity of final good production χ; and technique effects that lower $\Omega e(\theta)$ directly.

We examine balanced growth paths and impose two requirements on the set of paths we investigate. First, as usual we require non-deteriorating environmental quality. Second, we require positive growth in per capita income.[48]

To solve for a balanced growth path note Y/K must be constant along any such path. Using this requirement we can log differentiate (38) with respect to time, impose the requirement that Y/K be constant, and find the growth rate of final output:[49]

$$G_Y = (g+n) - \left[\frac{a_3}{1-a_1}\right][(g+n) - g_R] \tag{41}$$

The first term is the usual growth rate of output in the Solow model; the second is a negative element capturing the growth drag caused by natural resources. To see why this term appears suppose resources were in unlimited supply; then their services could grow over time at the same rate as effective labor and we would have $g_R = g + n$. In this case, capital, output, resources and effective labor would grow at the rate g+n and there would be no resource drag. In fact, however, the resource base $S(0) > 0$ is finite and exhaustible, and this implies that $g_R < 0$.[50] Any non-positive g_R is feasible because we can always choose the

[47] Algebra shows $a_1 = b_1 + b_2 b_3$, $a_2 = n_1 + b_2 n_2$, and $a_3 = b_2 b_4$.

[48] To this the reader may choose to add various efficiency conditions. For example, Stiglitz (1974) imposes the arbitrage condition requiring the return on capital equal that of the resource. This additional constraint is the well-known Hotelling (1931) result that the rate of capital gain on the resource in situ must equal the return on capital. This efficiency condition fails here because energy use creates the disutility pollution. We could impose a similar efficiency condition but its form would depend on how pollution entered utility. Alternatively, or in addition, the reader may add a condition requiring the marginal rate of substitution between consumption and pollution equal its marginal product. We leave a discussion of optimality until section 5.

[49] It is helpful to recall θ is constant and (38) is CRS.

[50] >From our stock equation we have $S(t) = S(0) - \int_o^t R(\tau)\, d\tau$ where $S(t)$ must be non-negative. This implies that $R(\tau)$ cannot rise over time along a balanced growth path.

level of resource use such that the finite stock is eliminated asymptotically. Therefore, the ratio of resources to effective labor in production falls over time and this reduces growth below the Solow level. Indeed as (41) shows, growth in final output could be negative if resource constraints loom too large.[51]

>From our earlier equations it is straightforward to show that $a_3 = b_2 b_4$ and hence the existence of finite resources lowers growth by an extent determined by the resource share in final output. To see this note that if the share of final output going to resources approached zero, then a_3 approaches zero, and G_Y in (41) approaches $g + n$ the Solow growth rate.

It is now straightforward to write growth in per capita output as just:

$$g_y = g - \left[\frac{a_3}{1 - a_1}\right][g + n - g_R] \tag{42}$$

which shows technological progress has to offset both population growth and the reduction in resources over time in order for per capita income to rise. To make this clear and relate our model here to the Stokey Alternative, suppose the resource in question offered an indestructible flow of services per unit time; i.e. suppose it was Ricardian land. Then $g_R = 0$ and growth in per capita income is positive if and only if:

$$a_2 g > a_3 n \tag{43}$$

The left hand side of (43) represents the Solow forces of technological progress the strength of which depends on the share of labor in overall production and the rate of labor augmenting technological progress. Aligned against these are the Malthusian forces lowering output per person by applying more and more labor to the fixed stock of land. The rate of population growth and the share of land in production determine the strength of the Malthusian forces. Note the similarity between (43) and our earlier (30). The condition in (43) arises when $g_R = 0$ and $g_y > 0$; the condition in (30) has $g_E = 0$, and $g_y > 0$. Note the parallel between emissions and resources.

To generate falling pollution we will need a strong compositional shift and hence $g_R < 0$ is our standard case. Our first condition for sustainability is that (42) is positive. Our second condition is that pollution must fall over time. Log differentiating our emissions function in (40) yields:

$$g_E = -g_A + g_\chi + g_Y + g_{e(\theta)} \tag{44}$$

to eliminate the possibility of emissions falling because of technological progress in abatement as in the Green Solow model, we set g_A to zero. To eliminate the possibility of greater abatement efforts holding pollution in check as in the Stokey Alternative, we set $g_{e(\theta)} = 0$. This leaves only changes in the composition of inputs to offset the rising scale effect of ongoing growth.

Straightforward calculations then show that the growth rate of energy per unit of final output is simply given by:

$$g_\chi = -[1 - b_3]\left[\left[\frac{b_4}{1 - b_3}\right] - \left[\frac{a_3}{1 - a_1}\right]\right][g + n - g_R] < 0 \tag{45}$$

The sign of (45) is negative (see footnote 47). Not surprisingly, the energy intensity of final output must fall over time.

Putting the growth rate of output and energy intensity together we find emissions will fall if and only if:

$$g_E = (g + n) - \left[b_3\left[\frac{a_3}{1 - a_3}\right] + b_4\right][(g + n) - g_R] < 0 \tag{46}$$

Note the first element in (46), $(g + n)$, is exactly the scale effect of output growth in the Green Solow model. And instead of technological progress in abatement appearing to offset the scale effect of growth, we

[51] Recall however the Solow result [See Solow (1974)] that with zero technological progress and zero population growth, a program of constant consumption can be maintained as long as the share of resources in final output is less than the share of capital. We cannot derive this condition from (41) directly because we have already imposed a constant Y/K, which is inconsistent with this program.

now have emissions per unit of final output falling from the composition effect given by the second negative term. The growth rate of output is lowered by the drag of natural resources and hence as the economy "abates" by altering its input mix this creates drag much as in Stokey (1998).

There is a tension therefore between the desirability of moving away from natural resource use in order to lower pollution emissions and the cost of doing so in terms of growth. This of course is a primary concern of many developing countries and has limited their participation in the Kyoto Protocol to limit global warming. Because of our addition of a fixed natural resource, in some cases, composition effects alone cannot generate positive balanced growth with a non-deteriorating environment.

To investigate we graph Gy from (42) and G_E from (46) in Figure 11 below. We have graphed these growth rates against the rate of change in effective labor per unit resource; that is against, $[(g+n)-g_R]$ since this term plays a key role in both emissions growth and per capita growth. There are several things to note about it. First, suppose resources were in unlimited supply; then their services could "grow" over time at the same rate as effective labor: $g+n$ and we would have $g_R = g+n$. Such a wonderful existence corresponds to points along the vertical axis in the figure. In particular we see that with no resource drag, per capita output growth equals g. With a finite resource base the growth rate of resource use must be negative and this means that per capita income growth must be lower as shown by the negatively sloped line labeled G_y starting at g and intersecting the horizontal axis at point B. Movements along this line correspond to changes in the growth rate of resource use g_R.

Similarly, if there were unlimited resources the energy intensity of GDP would remain constant and emissions would rise lock step with output. This unlimited resources scenario corresponds to a point along the vertical axis with the rate of aggregate output growth and emissions of $(n+g)$. Again, since resource use must decline over time the true growth rate of emissions must be lower as shown by the line labeled G_E that intersects the horizontal axis at A. The growth rate of emissions falls as we move to the right along this line because final output grows more slowly, and final output uses less energy per unit output.

>From these observations it is clear that at all points to the left of A, growth in emissions is positive; points to the right of A, growth in emissions is negative. Similarly, all points to the left of B have positive per capita output growth; points to the right have negative growth. Putting these results together we see that ongoing growth in per capita incomes and an improving environment may not be feasible in some cases. In particular, the bold line segment AB represents the feasible region. Taking g and n as exogenous, this region gives us a range of resource exploitation rates, g_R, that are consistent with our twin goals.[52]

To understand the determinants of the feasible region it proves useful to consider the zero population growth case. If population growth is zero, then the two lines have the same vertical intercept as shown by the dotted $n=0$ line that is parallel to G_E. Whether positive growth and falling emissions is possible only depends on the relative slopes of G_E versus G_y. Algebra tells us a region such as AB will always exist with zero population growth. The logic is simply that emissions growth falls with both reduced output growth *and* a changing energy intensity of production. Both occur as we increase drag by moving right in the figure. Therefore, once resource drag has lowered per capita output growth to zero at a point like B the scale effect is zero, but emissions growth must be strictly negative because the composition effect is still driving energy intensity downwards. Consequently, a feasible region like AB exists.

When population growth is positive, this logic fails. As we raise the population growth rate the G_E curve shifts to the right and eventually intersects the horizontal axis at B. This in effect raises the scale effect. At this point, positive growth with declining emissions is not possible. The reason is simply that emissions growth is rising in n (a scale effect), whereas growth in output per capita falls with n because of resource drag. Once we choose n large enough – as shown by the dashed line labeled $n_1 > n$ – the feasible region disappears.[53]

These results have a decidedly negative flavor to them. An environmental policy that lowers the growth rate of emissions and lowers the energy intensity of final output, also lowers per capita growth to such an

[52] You can derive the exact extraction level associated with any rate of exploitation by employing the materials balance constraint for resources.

[53] The issue of feasibility also arises in the Stokey Alternative although we didn't focus on it there. Recall Stokey (1998) assumed $n=0$. Our analysis here suggests that this is not an innocuous assumption.

extent that an improving environment and real income gains may be unattainable. There are several reasons why we should be cautious in interpreting these negative results. The first is simply that we have ruled out a role for active abatement as in the Stokey Alternative. And we have ruled out technological progress as in the Green Solow model. While adding more avenues of adjustment is always good, active abatement lowers pollution emissions but creates drag just as reducing energy use does. Routine calculations show that if we let all three avenues of adjustment operate, we can write our two balanced growth path requirements as:

$$G_y = \underbrace{g}_{\text{Green Solow}} - \underbrace{RD\left[(g+n)-g_R\right]}_{\text{natural resource drag}} + \underbrace{PPD\left[g_{e(\theta)}\right]}_{\text{pollution policy drag}} > 0 \tag{47}$$

$$G_E = \underbrace{g+n-g_A}_{\text{Green Solow}} - \underbrace{EI\left[(g+n)-g_R\right]}_{\text{composition effect}} + \underbrace{TE\left[g_{e(\theta)}\right]}_{\text{technique effect}} < 0$$

where RD is a positive constant representing resource drag, and PPD a positive constant representing pollution policy drag. Note that in general with both resource exhaustion and abatement rising, there are two sources of drag on per capita income growth. Corresponding to each source of drag is of course a component of emission reduction. In the second equation EI is a positive coefficient representing energy intensity changes. This corresponds to a composition effect. In addition TE is a positive coefficient representing changes coming from increased abatement; this represents a technique effect.

Putting all this together in terms of our figure, we find that allowing for technological progress in abatement ($g_A > 0$) shifts the growth of emissions line G_E inward expanding the feasible region. This should come as no surprise. Adding active abatement shifts both lines down (the economy grows slower as do emissions), having an ambiguous effect on the feasible region. Raising population growth from zero however shrinks the region making it more likely that both requirements cannot be met.

What then are we to make of our stylized facts from the introduction? Emission levels have been falling in many countries while growth in per capita income remained positive. Pollution abatement costs have trended upwards but only slowly, and energy prices – while rising – have not been rising at fast rates.[54] We have already seen that these features are roughly consistent with the Green Solow model and less so with the Stokey Alternative. Here we find that relying on changes in energy intensity alone can work in lowering emissions but it does so only with strong compositional shifts towards less energy intensive goods. In our formulation these shifts are only consistent with a rising real price of energy over time. To see this note that energy's share in final output is fixed; take final output as the numeraire, and conclude that the real price of energy must rise along the balanced growth path at the rate $-\chi > 0$.

In Figure 10 we plot the real price of three energy sources: oil, natural gas, and coal. For ease of reading all prices are set to 100 in 1957. It is very risky to draw any strong conclusions from this data. The real price of oil has almost doubled since 1957; the price of natural gas is rising quite quickly, while the price of coal has increased the least over the period. Naturally these price increases have created some composition effects as predicted by our source and sink model, but only over certain periods of time. For example, Wing and Ekhaus (2003) examine the history of energy intensity in US production and divide its changes into those accruing from a changing mix of US industries and those accruing from within industry improvements in energy efficiency (which would correspond to a fall in Ω). Their findings suggest that from the late 1950s until the mid 1970s changes in the composition of US industries played a major role in reducing overall energy intensity. But during the 1980s and 1990s the reduction in US aggregate energy intensity has come from improvements in energy efficiency at the industry level. Therefore changes in the composition of output cannot carry the burden of explanation of our data.

Instead these compositional changes must have been helped along by significant technological progress in abatement or energy efficiency (Ω). The evidence for these changes is very strong. For example in a detailed study of the energy efficiency of consumer durables Newell et al. (1999) find strong support for a significant role for autonomous technological progress (over 60% of the change in energy efficiency), and supporting

[54] Note from (37) that the relative price of energy to final good output must rise along the model's balanced growth path because we have already shown that the energy intensity of production, χ, falls over time (see (45)).

roles for induced innovation created by higher energy prices. Similar evidence is presented by Popp (2002) who examines the impact of higher energy prices on the rate of innovation in key energy technologies. Using a database of US patenting activity over the 1970-1993 period, Popp explains variation in the intensity of energy patenting across technology groups as a function of energy prices, the existing "knowledge stock" in a technology area and other covariates such as federal funding for R&D. There are two main results from the study. The first is that a rise in energy prices shown in Figure 10 created induced innovation and a burst in patenting activity after the oil price shocks.

The second major result is that while prices are a significant determinant of patenting activity, other factors are also very important. For example, the existing stock of knowledge (as measured by an index of previous patenting weighted for impact) in a technology area has a large impact on subsequent patenting. For example, Popp reports that the average change in knowledge stocks over the period raise patenting activity on average by 24%; while the average change in energy prices over the period raise patenting on average by only 2%. Knowledge accumulation and spillovers are very important in determining the pace of future innovation.[55]

Taking these considerations into account would likely expand our feasible region AB. For example, if the emission intensity Ω fell when either energy prices rose (as in the source-and-sink model) or abatement intensified (as in the Stokey Alternative), then composition changes and active abatement could play a smaller role in checking the growth of pollution. This would of course make feasibility more likely.

Adding complications to our existing models would however take us too far afield, and as yet we know of no research that explicitly links energy prices, induced innovation and pollution emissions within a growth framework. Instead we take a small step towards a theory of induced innovation in the next section when we introduce a model with learning by doing in abatement and reconsider our stylized facts. But before doing so, we should note that we have, to a certain extent, stacked the decks against sustainable growth by assuming environmental quality has no effect on production possibilities. We have assumed that reducing the flow of emissions has only a cost in terms of drag and no benefit in terms of heightened productivity in goods production due to higher environmental quality. Several authors have however postulated a direct and positive link between the productivity of final goods output and environmental quality. This link casts doubt on the validity of growth drag exercises like ours. A typical formulation would add to our models a shift term on the final goods production function that is increasing in environmental quality. For example, Bovenberg et al. (1995) and Tahovnen (1991) both postulate this type of additional interaction. Once we allow for a direct productivity response to an improved environment it is not clear that emission reductions lower growth. Bovenberg et al. and Tahovenen et al. both give sufficient conditions under which this additional channel dominates.

In general the less important are emissions in the direct production function, the more responsive is natural growth to a reduction in emissions, and the greater is the marginal productivity boost from a cleaner environment, the more likely it is that emission reduction could, in theory, boost growth. While it is certainly plausible that a deteriorating environment will lower productivity, it is however unclear how important these impacts are empirically. We suspect that for most of industrial production these environmental impacts are small. Certain industries such as farming or fishing are certain to have larger productivity effects from an improved environment, but these industries are small contributors to GDP in developed countries. It is likely that these induced productivity effects are greatest in poor developing economies and as yet have escaped the notice of serious empirical researchers.

While it is certainly possible for these direct productivity effects to exist, we feel the biggest restriction imposed by our analysis thus far is its failure to link the rising costs of pollution control to innovation targeted at raising the productivity of abatement. Induced changes in technology of this sort are likely to lower energy intensity over time given the price paths shown in Figure 10 and innovation in abatement

[55] Related empirical work by Kaufmann (2004) however is less sanguine about the ability of technical change to lower energy intensity in the long run. Kaufman examines the 1929-1999 period and argues that estimates of autonomous energy efficiency increases have been drastically overstated because changes in the composition of inputs and outputs have had led to a significant lowering of energy intensity. Instead he argues that inter fuel substitutions and reductions in household energy purchases are largely responsible for the declining trend in the energy intensity of GDP.

technologies is likely to be forthcoming as abatement costs rise. Both of these induced effects would lower emissions per unit final output by altering Ω. There is of course a large body of empirical research finding just such effects. But clearly these links are important although difficult to model in a growth framework, for as Popp notes

> "The most significant result [sic of the study] is the strong, positive impact energy prices have on new innovations. This finding suggests that environmental taxes and regulations not only reduce pollution by shifting behavior away from polluting activities but also encourage the development of new technologies that make pollution control less costly in the long run. . . ..simply relying on technological change is not enough. There must be some mechanism in place that encourages new innovation", p178.

With this quote in mind we now turn to consider the role of innovation induced by regulation.

5 Induced Innovation and Learning by Doing

The series of models we have examined thus far explain the growth and environment data by focusing on technological progress in goods production, increased abatement efforts or changes in the composition of output over time.[56] Missing from this list is a consideration of induced innovation lowering abatement costs. Induced innovation or learning by doing is prominent in both growth theory (since the writings of Arrow (1962)), and in environmental economics more generally. For example, Jaffe, et al (1995) stresses the role of induced technological advance in solving pollution problems and holding down abatement costs. New growth theory often adopts formulations that are in essence learning by doing models. In models where knowledge accumulates over time, innovators learn from this stock of knowledge. In models of human capital acquisition the evolution of human capital reflects the learning of past generations. And the simplest AK specification can be thought of a model where learning by doing in capital accumulation generates constant returns to capital accumulation at the economy wide level.

The introduction of learning by doing offers several new features to the growth and environment relationship. First, if abatement efforts are subject to learning by doing then this feature alone may generate the prediction of a first worsening and then improving environment. In a static setting, learning by doing is identical to increasing returns, and Andreoni and Levinson (2001) show how increasing returns to abatement can generate an EKC in a partial equilibrium endowment economy.[57]

Secondly, learning by doing alters the costs of pollution control. If learning by doing effects are unbounded, then growth with falling pollution levels could conceivably come at decreasing cost to society. In a world with bounded learning by doing the implications are less clear, but it seems likely that the drag of pollution policy would be smaller if learning by doing effects are present. An important feature of the static analysis mentioned above is that the authors generate falling pollution levels under quite weak assumptions on preferences. Specifically they do not need to adopt formulations where the demand for environmental protection is very income elastic. This suggests that a parallel dynamic analysis may escape these restrictions as well, because the cost of environmental control is now lower.

Third, if learning by doing arises from economy wide growth in the knowledge stock then learning by doing models offer the possibility of linking technological progress in abatement with that in goods production. Learning by doing models give us one way to make our assumptions about knowledge spillovers and technological progress consistent across sectors. And as our previous analysis makes clear, the relative rates of technological progress in goods production and abatement are key to determining the sustainability of a balanced growth path.

[56] To this we could add the set of papers invoking political economy arguments. See for example Jones and Manuelli (2001) and the related empirical work by Barrett (1998).

[57] Copeland and Taylor (2003) extend their analysis to a two-sector general equilibrium model with industry wide external economies in abatement and replicate their finding for a production economy.

Finally, although learning by doing is often modeled as a passive activity and not purposeful investment, learning by doing can be a form of induced innovation. If a worsening environment necessitates the imposition of pollution controls, and abatement is subject to learning by doing, we have effectively followed the advice of Popp in identifying a causal factor behind subsequent improvements in abatement technology.

5.1 Induced Innovation and the Kindergarten Rule Model

To discuss these issues, we now introduce the Kindergarten Rule model of Brock and Taylor (2003). This model, like those in the static literature, relies on learning by doing in the abatement process to hold pollution in check. Importantly, though since learning by doing is really an assumption about knowledge spillovers, the Kindergarten model adopts a consistent set of assumptions regarding the beneficial impact of knowledge spillovers. It assumes, similar to contributions in the AK growth literature, that knowledge spillovers in capital accumulation lead to constant returns at the aggregate level. Similarly, knowledge spillovers in abatement eliminate diminishing returns to abatement. As a consequence, we obtain a relatively simple model of growth with pollution controls where learning by doing reduces abatement costs but does not eliminate the drag of environmental policy entirely.

In order to focus on the *implications* of ongoing technological progress for the environment and growth, Brock and Taylor (2003) adopt the very direct link between factor accumulation and technological progress employed by Romer (1986), Lucas (1988) and others. By doing so they generate a simple one-sector model of endogenous growth and environmental quality.[58]

For simplicity they adopt a conventional infinitely lived representative agent, and assume all pollution is local. There is one aggregate good, labeled Y, which is either consumed or used for investment or abatement. There are two factors of production: labor and capital. There is zero population growth and hence $L(t) = L$; recall it is the rate of population growth relative to the rate of technological progress that is key, so here one of these rates is set to zero. In contrast to labor, the capital stock accumulates via investment and depreciates at the constant rate δ.

5.1.1 Tastes

A representative consumer maximizes lifetime utility given by:

$$W = \int_0^\infty U(C, X)\, e^{-\rho t} dt \tag{48}$$

where C indicates consumption, and X is the pollution stock. Utility is increasing and quasi-concave in C and $-X$ and hence $X = 0$ corresponds to a pristine environment. When we treat X as flow, $X = 0$ occurs with zero flow of pollution. When we allow pollution to accumulate in the biosphere we assume the (damaging) service flow is proportional to the level of the stock. Taking this factor of proportionality to be one, then X is the damaging service flow from the stock of pollution given by X.

A useful special case of $U(C, X)$ is the constant elasticity formulation:

$$\begin{aligned} U(C, X) &= \frac{C^{1-\varepsilon}}{1-\varepsilon} - \frac{BX^\gamma}{\gamma} \text{ for } \varepsilon \neq 1 \\ U(C, X) &= \ln C - \frac{BX^\gamma}{\gamma} \text{ for } \varepsilon = 1 \end{aligned} \tag{49}$$

where $\gamma \geq 0$, $\varepsilon \geq 0$ and B measures the impact of local pollution on a representative individual.

[58] Extensions of their framework to allow for purposeful innovation and a distinction between these two forms of capital, along the lines of Grossman and Helpman (1991) or Aghion and Howitt (1998), seem both feasible and worthwhile.

5.1.2 Technologies

The assumptions on production are standard. Each firm has access to a strictly concave and CRS production function linking labor and capital to output Y. The productivity of labor is augmented by a technology parameter T taken as given by individual agents. Following Romer (1986) and Lucas (1998) we assume the state of technology is proportional to an economy wide measure of activity. In Romer (1986) this aggregate measure is aggregate R&D, in Lucas (1988) it is average human capital levels; in AK specifications it is linked to either the aggregate capital stock or (to eliminate scale effects) average capital per worker. We assume T is proportional to the aggregate capital to labor ratio in the economy, K/L, and by choice of units take the proportionality constant to be one.[59]

5.1.3 From Individual to Aggregate Production

Although we adopt a social planning perspective, it is instructive to review how firm level magnitudes aggregate to economy wide measures since this makes clear the assumptions made regarding the role of knowledge spillovers. We aggregate across firms to obtain the AK aggregate production function as follows:[60]

$$\begin{aligned} Y_i &= F(K_i, TL_i) \\ Y &= \sum_i F(K_i, TL_i) \\ Y &= TLF(K/TL, 1) \\ Y &= KF(1,1) = AK \end{aligned} \tag{50}$$

where the first line gives firm level production; the second line sums across firms; the third uses linear homogeneity and exploits the fact that efficiency requires all firms adopt the same capital intensity. The last line follows from the definition of T.

Summarizing: diminishing returns at the firm level are undone by technological progress linked to aggregate capital intensity leaving the social marginal product of capital constant.

We now employ similar methods to generate the aggregate abatement technology. To start we note pollution is a joint product of output and we take this relationship to be proportional.[61] By choice of units we take the factor of proportionality to be one. Pollution emitted is equal to pollution created minus pollution abated. Abatement of pollution takes as inputs the flow of pollution, which is proportional to the gross flow of output Y^G, and abatement inputs denoted by Y^A. The abatement production function is standard: it is strictly concave and CRS. Therefore denoting pollution emitted by P, we can write pollution emitted by the ith firm as:

$$P_i = Y_i^G - a\left(Y_i^G, Y_i^A\right) \tag{51}$$

Now consider a Romeresque approach where individual abatement efforts provide knowledge spillovers useful to others abating in the economy. To do so we again introduce a technology shift parameter Γ, and assume it raises the marginal product of abatement. To be consistent with our earlier treatment of technological progress in production we assume Γ is proportional to the average abatement intensity in the economy, Y^A/Y^G. Then much as before we have the individual to aggregate abatement technology

[59] As is well known, one-sector models of endogenous growth blur the important distinction between physical capital and knowledge capital and force us to think of "capital" in very broad terms. Extensions of our framework along the lines of Grossman and Helpman (1991) or Aghion and Howitt (1998) seem both possible and worthwhile. These extensions would however add additional state variables making our examination of transition paths difficult.

[60] Implementing our planning solution by way of pollution taxes and subsidies to investment and abatement should be straightforward.

[61] Nothing is lost if we assume pollution is produced in proportion to the services of capital inputs. The service flow of capital is proportional to the stock of capital, and the stock of capital is proportional to output.

transformation as:

$$\begin{aligned} P_i &= Y_i^G - a\left(Y_i^G\Gamma, Y_i^A\right) \\ P_i &= Y_i^G\left[1 - \Gamma a\left(1, Y_i^A/Y_i^G\Gamma\right)\right] \\ \sum_i P_i &= \sum_I Y_i^G\left[1 - \Gamma a\left(1, Y_i^A/Y_i^G\Gamma\right)\right] \\ P &= Y^G\left[1 - \Gamma a\left(1,1\right)\right],\ a\left(1,1\right) > 1 \\ P &= Y^G\left[1 - \theta a\left(1,1\right)\right],\ \theta \equiv Y^A/Y^G \end{aligned} \tag{52}$$

where the first line introduces the technology parameter Γ; the second exploits linear homogeneity of the abatement production function; the third aggregates across firms; the fourth recognizes that efficiency requires all firms choose identical abatement intensities, uses the definition of Γ and notes that for abatement to be productive it must be able to clean up after itself. The fifth line defines the intensity of abatement, $\theta \equiv Y^A/Y^G$. Since abatement can only reduce the pollution flow we must have $\theta \leq 1/a\left(1,1\right)$.[62]

It is important to note that the aggregate relationship between pollution and abatement given by the last line in (52) is consistent with empirical estimates finding rising marginal abatement costs at the firm level. Each individual firm has abatement costs given by foregone output used in abatement, and hence partially differentiating the first line of (52) and rearranging we find:

$$\frac{\partial Y_i^A}{\partial P_i} = -1/\left[\frac{\partial a}{\partial Y_i^A}\right] < 0,\ \frac{\partial Y_i^{A2}}{\partial P_i^2} > 0 \tag{53}$$

Marginal abatement costs are rising at the firm level.

Marginal abatement costs at the society level, are however, constant. To see why totally differentiate (52) allowing Γ and individual abatement to both vary. We find:

$$\frac{dY_i^A}{dP_i} = -1/\left[\partial a/\partial Y_i^A\right] - \frac{\left[\partial a/\partial Y_i^G\right]}{\left[\partial a/\partial Y_i^A\right]}\frac{d\Gamma}{dP_i} \tag{54}$$

Γ is the average abatement intensity in the economy, which given identical firms, is just the abatement intensity for the typical ith firm. Using $\frac{d\Gamma}{dP_i} = \left[\frac{1}{Y_i^G}\right]\left[\frac{dY_i^A}{dP_i}\right]$ and rearranging we obtain:

$$\begin{aligned} \frac{dY_i^A}{dP_i} &= -\frac{Y_i^A}{\left[\left[\frac{\partial a}{\partial Y_i^A}\right]Y_i^A + \left[\frac{\partial a}{\partial Y_i^G}\right]\Gamma Y_i^G\right]} \\ &= -\frac{Y_i^A}{a\left(Y_i^A, \Gamma Y_i^G\right)} = -\frac{1}{a\left(1,1\right)} < 0 \end{aligned} \tag{55}$$

where the first line follows from rearrangement and the second by CRS in abatement. The result given in (55) is identical to what we find by differentiating the aggregate relationship between pollution and abatement given in the last line of (52).

Summarizing: diminishing returns at the firm level, that lead to rising marginal abatement costs, are undone by technological progress linked to aggregate abatement intensity leaving the social marginal cost of abatement constant.

The formulations of learning by doing that we have adopted are extreme. In general we would expect the productivity in abatement (or production) to adjust gradually in response to a slow moving measure of knowledge capital. In the cases developed here however the productivity boost from an increased knowledge

[62] Adding the possibility of investments in restoration would probably strengthen the case for sustainable growth. Abatement of pollution and restoration are however distinct activities. We imagine that a restoration production function would take as an input the current damage to the environment – our stock variable X – and then apply inputs to restore it. This is quite different from abatement which operates to lower the current flow of pollution by use of variable inputs.

capital occurs instantaneously. So instead of Γ being a complicated function of the abatement intensities adopted in the infinite past history of the economy weighed by their relevance to productivity today, it is simply proportional to the current intensity. This is of course an abstraction, but a useful one since it frees us from keeping track of the evolution of two additional state variables (knowledge capital in abatement and knowledge capital in production), and allow us to capture the main feature of learning by doing models by linking the productivity of abatement to the intensity of this activity at the economy wide level. It also yields simple linear forms for production and abatement that add greatly to the model's tractability. This last feature is especially important in a model where the stock of environmental quality has already raised the number of state variables to two.

Putting these pieces together our planner faces the aggregate production relations for output and abatement given by the last lines of (50) and (52) together with the atemporal resource constraint linking gross output, abatement and net production:

$$Y = Y^G - Y^A \tag{56}$$

The Kindergarten model is only one approach to modeling endogenous growth and environment interactions. Closely related approaches in an AK framework are those of Stokey (1998), Smulders (1993) and Smulders and Gradus (1996). These papers all adopt AK models, but end up with different conclusions. Early work in a one-sector framework by Smulders (1993) and Smulders and Gradus (1996) demonstrated how continuing economic growth and constant environmental quality are compatible in an AK model. In constrast, Stokey (1998) demonstrated how continuing growth and constant environmental quality are not compatible within a AK set up. The difference in their results comes from their different assumptions on abatement. To see why this is true, start with (51), ignore knowledge spillovers, and work forward using now familiar steps to find:

$$P_i = Y_i^G \phi\left(1-\theta\right), \ \phi\left(\theta\right) \equiv \left[1 - a\left(1, 1-\theta\right)\right] \tag{57}$$

Stokey employs the specific functional form for ϕ given by $\left(1-\theta\right)^{\beta}$ for $\beta > 1$, and this implies the CRS abatement production function given by:

$$a\left(Y_i^G, Y_i^A\right) = Y_i^G \left[1 - \left(1 - \frac{Y_i^A}{Y_i^G}\right)^{\beta}\right] \tag{58}$$

Using (57) it is now easy to show abatement is subject to diminishing returns:

$$\frac{\partial P_i}{\partial Y_i^A} = -\beta\left(1-\theta\right)^{\beta-1} < 0, \ \frac{\partial P_i^2}{\partial Y_i^{A2}} > 0$$

which implies marginal abatement costs are rising at the aggregate level. Setting $Y^A = 0$ we find the first unit of abatement lowers pollution by the amount $\beta > 1$, somewhat similar to our formulation where $a(1,1) > 1$.[63] If we now combine (57) with (50), recall net output is $(1-\theta)$ times gross output, and introduce the variable $z = 1-\theta$, we find the exact specification employed in Stokey (1998).

$$Y = AKz, \ P = AKz^{\beta} \tag{59}$$

Stokey's (1998) result that growth is not possible follows from matching an AK aggregate production function with strictly neoclassical assumptions on abatement adopted from Copeland and Taylor (1994). That is, if we think of the AK model as one of knowledge spillovers then Stokey has assumed these spillovers occur in production but not abatement. By doing so, she eliminates "technological progress" in abatement and this eliminates the possibility of sustainable growth.

Comparing our approach to the work of Smulders is more difficult because abatement is not specifically modeled and he considers a variety of formulations. By specializing his framework to the AK paradigm we find:

$$Y = \alpha K, \ P = \left[\frac{K}{A}\right]^{\gamma} \tag{60}$$

[63] Since the marginal product of abatement is bounded when abatement is zero, Stokey (1998) is able to show that no regulation is undertaken initially and pollution rises lock-step with output.

The first element is just a standard AK production function. The second relates what Smulders refers to as net or emitted pollution to the capital stock, K, and abatement, A. If we employ (60) and solve for emissions per unit of gross output we find:

$$\frac{P}{Y} = \left[\frac{K}{A}\right]^{\gamma} / \alpha K \tag{61}$$

If the economy allocates a fixed fraction of its output to abatement, K/A is constant, and emissions per unit of gross output fall with the size of the economy. This reflects a strong degree of increasing returns. Moreover, the reader may note from (60) that pollution emitted goes to infinity as abatement goes to zero, which is inconsistent with pollution being a joint product of output. Therefore, Smulders and Gradus (1993, 1996) match AK aggregate production with assumptions on abatement ensuring increasing returns; and, in contrast with the Kindergarten specification, assume pollution is not a joint product of output.

5.1.4 Endowments

We treat pollution as a flow that either dissipates instantaneously – such as noise pollution – or a stock that is only eliminated over time by natural regeneration – such as lead emissions or radioactive waste. When X is a stock we have:

$$\dot{X} = AK\left[1 - \theta a\right] - \eta X \tag{62}$$

where η represents the speed of natural regeneration, and where for economy of notation we have denoted $a(1,1)$ by a. When X is a flow we have:

$$X = AK\left[1 - \theta a\right]$$

5.1.5 The Kindergarten Rule

We focus first on the possibility of balanced and continual growth, leaving to the next section a discussion of transition paths. Before we proceed with the formal analysis it proves instructive to step back slightly to consider the feasibility and optimality of sustainable growth. From our assumptions on abatement it is clear that if θ is set high enough all pollution emissions will be eliminated and we will enter a zero emission world. Therefore as long as $a > 1$ there will exist a $\theta < 1$ that generates zero emission technologies. And if $\theta < 1$ then some output will be left over for consumption and investment which can in turn drive growth in output. It appears then that feasibility is guaranteed by knowledge spillovers in abatement generating a constant marginal product.

The assumption of $a > 1$ is innocuous. Recall that abatement, like all other economic activities, pollutes. One unit of abatement creates one unit of pollution, but cleans up $a > 1$ units of pollution. It is only this surplus between costs and benefits, $1 - 1/a > 0$ that makes abatement useful at all. But even if growth is feasible, abatement is costly and this will cause drag as in our earlier formulations. The remaining questions for sustainability are how large is this drag, how much will it lower the return to capital, and what restrictions on preferences will be needed to generate sustainable growth.

To answer these questions consider the following problem:

$$\begin{aligned} & Maximize \int_0^{\infty} U\left(C, X\right) e^{-\rho t} dt \\ \text{s.t.} \quad & K\left(0\right) = K_0, \; X\left(0\right) = X_0, \text{ and } \theta \leq 1/a \\ & \dot{K} = AK\left[1 - \theta\right] - \delta K - C \\ & \dot{X} = AK\left[1 - \theta a\right] - \eta X \end{aligned} \tag{63}$$

where we adopt $U(C, X)$ from (49). Recall the fraction of gross output allocated to abatement is θ and since the flow of pollution into the environment cannot be negative this will never exceed $1/a$. We can write the

Hamilton-Jacobi-Bellman equation as:

$$\rho W(K,X) = Max \left\{ \begin{array}{c} H = \frac{C^{1-\varepsilon}}{1-\varepsilon} - \frac{BX^{\gamma}}{\gamma} + \\ \\ \lambda_1 \left[AK\left[1-\theta\right] - \delta K - C \right] + \lambda_2 \left[AK\left[1-\theta a\right] - \eta X \right] \end{array} \right\} \tag{64}$$

where $\rho W(K,X)$ is the maximized value of the program for the given initial conditions $\{K_0, X_0\}$, and H is the current value Hamiltonian for our problem. The controls for this problem are consumption, C, and abatement intensity, θ.

Observe the term involving our control variable, θ,

$$Max\left\{AK\theta\left[-\lambda_1 - a\lambda_2\right]\right\} \text{ s.t. } 0 \leq \theta \leq 1/a$$

where λ_1 is the positive shadow value of capital and λ_2 is the negative shadow cost of pollution. Since the Hamiltonian is linear in θ, the value of the term $S = [-\lambda_1 - a\lambda_2]$ will largely determine the optimal level of abatement. When $S > 0$, the shadow cost of pollution is high relative to that of capital. In this case abatement is relatively cheap and maximal abatement will be undertaken. Conversely when $S < 0$ the shadow value of capital is high relative to that of pollution. In this case abatement is relatively expensive and zero abatement will occur. Finally, when $S = 0$, the shadow values are equated and active, but not necessarily maximal, abatement will occur. Therefore, the value of S determines when and if the economy switches from a zero-to-active-to-maximal abatement regime. We deal with these possibilities in turn.

Regardless of the value of S, the optimal level of consumption will always satisfy

$$\frac{\partial H}{\partial C} = C^{-\varepsilon} - \lambda_1 = 0 \tag{65}$$

although the shadow value of capital and its dynamic path may differ across regimes.

When $S > 0$, maximal abatement occurs and the dynamics are given by:

$$\begin{array}{rcl} S & > & 0 \\ \theta & = & \theta^K,\ \theta^K \equiv 1/a \\ \dot{\lambda}_1 & = & -g\lambda_1,\ g \equiv A\left[1-\theta^K\right] - \delta - \rho \\ \dot{K} & = & \left[g+\rho\right]K - C\left(\lambda_1\right),\ K\left(0\right) = K_0,\ C\left(\lambda_1\right) \equiv \lambda_1^{-1/\varepsilon} \\ \dot{\lambda}_2 & = & \lambda_2\left[\rho+\eta\right] + BX^{\gamma-1} \\ \dot{X} & = & -\eta X,\ X\left(0\right) = X_0 \end{array} \tag{66}$$

By choosing the intensity of abatement $\theta = \theta_K$ there are no net emissions of pollution and the environment improves at a rate given by natural regeneration. We dub θ_K "the Kindergarten rule" because when economies adopt the Kindergarten rule pollution is cleaned up when it is created.[64]

Alternatively, S may be exactly zero. In this case we have an interior solution for abatement, with the following dynamics:

$$\begin{array}{rcl} S & = & 0 \\ \theta & \in & \left[0, \theta^K\right] \\ \dot{\lambda}_1 & = & -g\lambda_1 \\ \dot{K} & = & \left[A\left[1-\theta\right] - \delta\right]K - C\left(\lambda_1\right),\ K\left(0\right) = K_0,\ C\left(\lambda_1\right) \equiv \lambda_1^{-1/\varepsilon} \\ \dot{\lambda}_2 & = & \lambda_2\left[\rho+\eta\right] + BX^{\gamma-1} \\ \dot{X} & = & AK\left[1-a\theta\right] - \eta X,\ X\left(0\right) = X_0 \end{array} \tag{67}$$

[64] This is one of the most common rules taught in Kindergarten. For a list of common Kindergarten rules see All I Really Need to Know I Learned in Kindergarten: Uncommon Thoughts on Common Things by Robert Fulgham. Fulgham argues that the basic values we learned in grade school such as "clean up your own mess" (in effect our Kindergarten rule) and "play fair" are the bedrock of a meaningful life.

In this situation pollution is not completely abated, and hence the evolution of environmental quality reflects both the level of active abatement and natural regeneration. And finally, with no abatement at all we must have $S < 0$. Both pollution and output rise over time yielding:

$$\begin{aligned} S &< 0 \\ \theta &\in 0 \\ \dot{\lambda}_1 &= -\lambda_1 \left[A - \delta - \rho\right] - \lambda_2 A \\ \dot{K} &= \left[A - \delta\right] K - C\left(\lambda_1\right),\ K\left(0\right) = K_0,\ C\left(\lambda_1\right) \equiv \lambda_1^{-1/\varepsilon} \\ \dot{\lambda}_2 &= \lambda_2 \left[\rho + \eta\right] + BX^{\gamma - 1} \\ \dot{X} &= AK - \eta X,\ X\left(0\right) = X_0 \end{aligned} \tag{68}$$

Consider growth paths with active abatement. Then from (67) and (66) we find the shadow value of capital falls over time at a constant exponential rate:

$$-g \equiv \frac{\dot{\lambda}_1}{\lambda_1} = -\left[A\left[1 - \theta^K\right] - \rho - \delta\right] < 0 \tag{69}$$

provided the net marginal product of capital, at the Kindergarten rule level of abatement, A[1-θ^K], can cover both depreciation and impatience. We leave for now a detailed discussion of what this requires and assume it is true: $g > 0$. Then it is immediate that consumption rises at the constant rate $g_C = g/\varepsilon > 0$.

>From the capital accumulation equations in both (66) and (67) we can now deduce that capital and output must grow at the same rate as consumption if θ is constant over time. To determine whether the intensity of abatement is constant over time, consider the accumulation equation for pollution:

$$\dot{X} = AK\left[1 - \theta a\right] - \eta X \tag{70}$$

There are two ways (70) can be consistent with balanced growth. The first possibility is that we have a maximal abatement regime where $S > 0$ holds everywhere along the balanced growth path. In this situation, K grows exponentially over time and θ is set to the Kindergarten rule level. Using (66), this balanced growth path must have:

$$\dot{X} = -\eta X \text{ and } \theta = \theta^K \tag{71}$$

In this scenario, the environment improves at the rate η over time and abatement is a constant fraction of output $1 > \theta^K > 0$. As time goes to infinity the economy approaches a pristine level of environmental quality. Therefore the balanced growth path exhibits constant growth in consumption, output, capital and environmental quality. Consumption is a constant fraction of output and we have:

$$g_c = g_k = g_y = g/\varepsilon > 0,\ g_x = -\eta < 0 \tag{72}$$

A second possibility is that abatement is active but not maximal. Define the deviation of abatement from the Kindergarten rule as $D(\theta) = (\theta^K - \theta)/\theta^K$. Using this definition rewrite (70) to find:

$$\frac{\dot{X}}{X} = \frac{AK\left[D\left(\theta\right)\right]}{X} - \eta \tag{73}$$

It is apparent that if the deviation of abatement from the Kindergarten rule fell exponentially, then it may be possible for X to fall exponentially while K rises. That is, in obvious notation, a possible balanced growth path would have:

$$g_k + g_D = g_x < 0 \tag{74}$$

In this situation abatement is at an interior solution at all times and becomes progressively tighter over time approaching the Kindergarten rule asymptotically. The inflow of pollution from production into the environment is always positive but environmental quality improves nevertheless. This intuitive description

suggests that the possibility of this outcome must rely on both the pace of economic growth and the ability of the environment to regenerate. This is indeed the case as Brock and Taylor (2003) show that a necessary condition for us to remain in a $S = 0$ regime with active abatement is simply:

$$\eta\left(\gamma - 1\right) > g \tag{75}$$

This condition reflects two different requirements. The first is simply that γ cannot equal one. If it did then the (instantaneous) marginal disutility of pollution is a constant and λ_2 is a constant as well. This would also imply that consumption be fixed as well. This is inconsistent with growth of any sort.

Assuming γ not equal to one is necessary for balanced growth with an interior solution for abatement. But a second condition must also hold. Natural regeneration, η, must be sufficiently large relative to the growth rate g. If the rate of regeneration is high and growth rates quite low, then the optimal plan is to use nature's regenerative abilities to partially offset the costs of abating because the shadow value of foregone output is high in slow growth situations. Conversely, if regeneration is low and the growth rate g relatively high, then no amount of abatement short of the Kindergarten rule will hold pollution to acceptable levels.

This intuition suggests a natural corollary for the case of flow pollutants. If pollution has only a flow cost it is "as if" the environment is regenerating itself infinitely fast. This intuition suggests that as we let η get large, the results in the stock pollutant case should replicate those for a flow. This intuition is, in fact, correct. Brock and Taylor prove that when $g > 0$ and X is a flow pollutant, then sustainable economic growth with an ever improving environment is possible and optimal. With a flow pollutant, if $\gamma > 1$, then the intensity of abatement approaches the Kindergarten rule level of abatement, θ^K, asymptotically. Alternatively, if $\gamma = 1$, then $\theta = \theta^K$ everywhere along the balanced growth path.

These results are important in showing how the Kindergarten rule generates sustainable growth. Sustainable growth requires two conditions. The first is that g given in (69) is positive. The assumption $g > 0$ requires the marginal product of capital, adjusted for the ongoing costs of abatement, be sufficiently high. A necessary condition is that $A[1 - \theta^K]$ be positive, but this is guaranteed as long as abatement is a productive activity. Given abatement is productive, we still require the adjusted marginal product of capital, $A[1 - \theta^K]$, to offset both impatience and depreciation. If abatement is not very productive, then $\theta^K = 1/a$ will be close to one and growth cannot occur. If capital is not very productive or if the level of impatience and depreciation are high then ongoing economic growth cannot occur. These are however very standard requirements for growth under any circumstances.[65]

The second is that $h = g(1 - 1/\varepsilon) + \rho > 0$. This condition is the standard sufficiency condition for the existence of an optimum path in an AK model with power utility.[66] This condition is of course weaker than that needed in earlier models generating declining pollution levels. For example, ε is just σ in the CRRA specification we used earlier and we have already seen that Stokey (1998), Lopez (1998) and others require $\sigma > 1$ to generate declining emissions. Here the requirement is far weaker and this follows from the fact that consumer's are not required to make larger and larger sacrifices in consumption to fund an every growing abatement program.

5.2 Empirical Implications

The Kindergarten model relies heavily on the assumed role of technological progress in staving off diminishing returns to both capital formation and abatement. It is impossible to know apriori whether technological progress can indeed be so successful and hence it is important to distinguish between two types of predictions before proceeding. The first class of predictions are those regarding behavior at or near the balanced growth

[65] In Keeler, Spence and Zeckhauser (1972) a similar condition describes their Golden Age capital stock. In their model with no endogenous growth the Golden Age capital stock is defined by (in our notation) the equality $f'(K)[1 - 1/a] - \delta = \rho$ simulations of the model assume a to be 12 (see p.22). Chimeli and Braden (2002) assume a similar condition. Both studies assume abatement or clean up is a linear function of effort thereby ignoring the reality of diminishing returns and the necessity of ongoing technological progress.

[66] Denote the growth rate in an AK model with power utility by g^*, then in terms of our parameters we have $g^* = g/\varepsilon$ and the standard condition is $\rho + (\varepsilon - 1)\, g^*$. This is equivalent to $h > 0$. See Aghion and Howitt (1998, Equation (5.3)).

path. This set has received little attention in the empirical literature on the environment and growth, although balanced growth path predictions and their testing are at the core of empirical research in growth theory proper (see the review by Durlauf and Quah (1999)). The second set of predictions concern the transition from inactive to active abatement and these are related to the empirical work on the Environmental Kuznets Curve (See Grossman and Krueger (1993,1995) and the review by Barbier (2000)).

5.3 Balanced Growth Path Predictions

Using our previous results it is straightforward to show that near the balanced growth path we must have: convergence in the quality of the environment across all countries sharing parameter values but differing in initial conditions; the share of pollution abatement costs in output approaching a positive constant less than one; overall emissions rates falling and environmental quality rising; and emissions per unit output falling as production processes adopt methods that approach zero emission technologies. The model also presents predictions for the intensity of abatement that we discuss subsequently.

Whether the cross-country predictions will be borne out by empirical work is as yet unknown but an examination of US data shows the model's strongest predictions – those regarding falling emissions and improving environmental quality - are not grossly at odds with available U.S. data. The most favorable evidence for the model is the slow movement in pollution abatement costs in the face of dramatically declining pollution levels. The model explains this feature of the data by recourse to specifics of the abatement function that hold abatement costs down much as exogenous technological progress does in the Green Solow model.

The prediction of declining emission intensities along the balanced growth path is also consistent with the data shown in Figure 1, but as in Stokey (1998) the model only predicts declining emissions to output ratios after regulation begins.

5.4 The Environmental Catch-up Hypothesis

We have so far focused on balanced growth paths but the large EKC literature concerns itself with what must be transition paths towards some BGP. To examine these predictions we present several transition paths in Figure 5. One of these paths is that of a Poor country having small initial capital K^P but a pristine environment. The other is the path of a Rich country starting again with a pristine environment but with a much larger initial capital K^R. These differences in initial conditions could reflect variance across countries in geography, resource endowments or institutions that impact on initial productivity levels. Each economy starts with a pristine environment in stage I and grows. During this stage there is no pollution regulation: the environment deteriorates, X rises, and the capital stock grows until the trajectory hits the Switching Locus labeled SL. Once the economy hits the Switching Locus active regulation begins and the economy enters stage II.[67]

It is apparent from the figure that the Poor country experiences the greatest environmental degradation at its peak, and at any given capital stock, (i.e. income level) the initially Poor country has worse environmental quality than the Rich. Moreover, since both Rich and Poor economies start with pristine environments, the qualities of their environments at first diverge and then converge. This is the Environmental Catch-up Hypothesis.

Divergence occurs because the opportunity cost of abatement (and consumption) is much higher in capital poor countries. A high shadow price of capital leads to less consumption, more investment and rapid industrialization in the Poor country. Nature's ability to regenerate is overwhelmed. The quality of the environment falls precipitously. In capital rich countries the opportunity cost of capital is lower: consumption is greater and investment less. Industrialization is less rapid and natural regeneration has time to work. The peak level of environmental degradation in the Rich country is therefore much smaller. But

[67] Brock and Taylor (2003) show the exact position and shape of the locus depend on whether parameters satisfy the fast growth or slow growth scenario. For the most part we will proceed under the assumption that economic growth is fast relative to environmental regeneration; that is (75) fails strongly and we have $\eta(\gamma - 1) < g$. This implies $\theta(t) = \theta^K$ everywhere along the balanced growth path (Figure 12 implicitly assumes this result). For illustrative purposes we will sometimes discuss the parallel flow case (where we can think of η approaching infinity but (75) failing because $\gamma = 1$).

once we enter Stage II abatement is undertaken and since abatement is an investment in improving the environment, it is only undertaken when the rate of return on this investment equals (or exceeds) the rate of return on capital. Since economies are identical, except for initial conditions, rates of return are the same across all countries in Stage II. Equalized rates of return require equal percentage reductions in the pollution stock. Therefore absolute differences in environmental quality present at the beginning of Stage II disappear over time.

Note how similar this intuition for the ECH is to that given for the EKC prediction in the Green Solow model. In the Green Solow model initial rapid growth overwhelms nature's ability to dissipate pollution starting from its initial position at a biological equilibrium. Eventually growth slows and the environment's regenerative powers restore its quality slowly over time. Growth is initially rapid in the Solow model because of diminishing returns. In the Kindergarten model, growth is initially rapid because there is no regulation and no drag from pollution policy to lower the marginal product of capital. And once regulation is active, growth slows because regulation lowers the net marginal product of capital. The environment's regenerative powers then restore its quality slowly over time. Both explanations have nature overwhelmed early on and both give prominent roles to a declining marginal product of capital.

The discussions above, and Figure 5, assume the fast-growth-slow-regeneration assumptions hold. We chose this case to discuss and illustrate because it illustrates the forces at work very clearly. Since many of the same conclusions hold when growth is relatively slow we only provide a sketch here of some differences. There is again a Switching Locus which divides Stage I from Stage II. The Switching Locus again defines a unique X^* that is declining in K^*. The most important difference is that once a trajectory of the system hits this new Switching Locus it remains within it forever. If the economy is below the locus then abatement is inactive and K rises at a rapid rate: X increases rapidly until the Switching Locus is reached in finite time. If the initial (K, X) is above the locus, maximal abatement is undertaken but this drives down the shadow cost of pollution very quickly and we again hit the locus, this time, from above.

Once on the locus, countries remain trapped within it thereafter and this implies the economy's choice of abatement remains interior; i.e. the trajectory follows along the Switching Locus maintaining $MAC = MD(K, X)$ throughout. Over time abatement rises and the intensity of abatement approaches the Kindergarten rule in the limit. Therefore, the slow growth case is very similar except that the model now predicts an even stronger form of convergence. All transition paths remain on the Switching Locus once active abatement begins; therefore policy active countries share the *same* path for environmental quality and income levels in Stage II.

5.5 The ECH and the EKC

Brock and Taylor prove that all economies capable of sustained growth must follow the stage I – stage II life cycle producing an EKC like relationship between income and environmental quality.[68] Their income and growth prediction is however somewhat different from a standard EKC result. They predict that countries differentiated only by initial conditions exhibit initial divergence in environmental quality followed by eventual convergence.[69] Moreover, as Figure 5 makes clear countries make the transition to active abatement at different income and peak pollution levels. This of course throws into question empirical methods seeking to estimate a unique income-pollution path. More constructively it suggests that an important feature of the data may well be a large variance in environmental quality at relatively low-income levels with little variance at high incomes. Empirical work by Carson et al. (1997) relating air toxics to U.S. state income levels is supportive of this conjecture:

[68] The addition of perfect capital markets can affect the pollution and income path greatly. To a certain extent a country is running down its environment initially to accumulate capital. With *perfect* capital markets it is possible to eliminate stage I entirely in some cases. Given many less developed countries have very limited and imperfect access to capital markets it is difficult to know the empirical importance of this result. It suggests however that access to capital markets may be an excluded country characteristic in EKC style regressions.

[69] It is possible to show X is rising throughout Stage I and this ensures points along the Switching Locus do indeed represent peaks in pollution levels.

> "Without exception, the high-income states have low per-capita emissions while emissions in the lower-income states are highly variable. We believe that this may be the most interesting feature of the data to explore in future work. It suggests that it may be difficult to predict emission levels for countries just starting to enter the phase, where per capita emissions are decreasing with income", p. 447-8.

In some cases however, (initially) Rich and Poor will make the transition at the same income level but still exhibit our Catch-up Hypothesis. To investigate, we report the switching locus in the fast growth case. It takes on an especially simple form given by:

$$MAC = \frac{1}{a} = \frac{B}{[\rho + \gamma\eta]} X^{*\gamma-1} [hK^*]^{\varepsilon} = MD(K^*, X^*) \tag{76}$$

which is the downward sloping and convex relationship between pollution and capital depicted in Figure 5. The left hand side of (76) represents marginal abatement costs. The right hand side is marginal damage evaluated at $\{K^*, X^*\}$. Marginal damage is increasing in the pollution stock provided γ exceeds one, and since the flow of national (and per-capita) income Y is always proportional to K, it is apparent that the income elasticity of marginal damage with respect to flow income is given by ε. Large values of ε correspond to the strong income effects referred to earlier

Let γ approach one. Then the slope of the Switching Locus approaches infinity and all countries attain their peak pollution levels at the same K^*. But even with a common turning point differences in environmental quality remain. Moreover, these are not simple level differences because countries initially diverge and then converge after crossing K^*. To eliminate our catch-up hypothesis we must assume regeneration is infinitely fast: X is a flow. In this case Brock and Taylor show the Switching Locus is again vertical at a given K^*. More importantly, since pollution is proportional to production before K^*, and policies are identical after K^*: initial conditions no longer matter.

These results tell us that when pollution is strictly a flow, all countries share the same income pollution path. But when pollution does not dissipate instantaneously, initial conditions matter. We have the Environmental Catch-up Hypothesis, and empirical methods must now account for the persistent role of initial conditions.[70]

5.6 Pollution Characteristics

The ECH focuses on cross-country comparisons in pollution levels, but says little about how predictions vary with pollutant characteristics. And while most authors have focused on generating an EKC relationship there has been very little work examining how these predictions vary with pollutant characteristics. This is unfortunate since there is good data in the U.S. and elsewhere that could be fruitfully employed to test within country but across pollutant predictions. This is especially important since many models can generate the EKC result.

To demonstrate the Kindergarten model's across pollutant predictions consider regeneration first and start from a position where $\eta = 0$ (radioactive waste). Brock and Taylor (2003) show that the Switching Locus in Figure 5 shifts outwards as we raise η. The response is to delay action and allow the environment to deteriorate further. Once we raise η sufficiently the economy eventually enters the fast regeneration regime and here we find abatement delayed in another manner – it is introduced slowly by the now gradual implementation of the Kindergarten rule. Faster regeneration then implies that countries either begin abatement at higher income levels or allow their environments to deteriorate more before taking action. Surprisingly, fast regeneration will be associated with lower and not higher environmental quality - at least over some periods of time or ranges of income.[71]

[70] This result may explain why empirical research investigating the EKC has been far more successful with air pollutants like SO2, than with water pollutants or other long lasting stocks (see the review by Barbier (2000)).

[71] An especially colorful example of delay in abatement caused by rapid natural regeneration is that of the City of Victoria in British Columbia. Every day, the Victoria Capital Regional District (CRD) dumps approximately 100,000 cubic meters

A change in regeneration rates also affects the pace of abatement. If a pollutant has a long life in the environment, then once abatement begins it is clear that natural regeneration can play only a small role. Consequently the optimal plan calls for an initial period of inaction before starting a very aggressive abatement regime: the immediate adoption of the Kindergarten rule. When η is relatively large we are in the fast regeneration regime and abatement is intensified gradually and only approaches the Kindergarten rule level in the limit.

Putting the predictions for the timing and intensity of abatement together, Brock and Taylor find that very long-lived pollutants should be addressed early with their complete elimination compressed in time. It is optimal to delay action on short-lived pollutants and adopt only a gradual program of abatement. This description of optimal behavior is of course consonant with the historical record in several instances where long-lived chemical discharges and gas emissions were eliminated very quickly by legislation, whereas short-lived criteria pollutants have seen active regulation but not elimination over the last 30 years.

Pollutants also differ in their toxicity. The marginal disutility of toxics could exceed those classified as irritants, and damages from toxics may rise more steeply with exposure. The first feature of toxics implies their abatement should come early. This is clear from (76) where increases in B shift the Switching locus inwards and hasten abatement. Surprisingly very convex marginal damages (a high γ) call for the gradual and not aggressive elimination of pollution. The logic is that any reduction in the concentration of toxics has a large impact on marginal damage. Therefore, only by lowering emissions slowly can we match a steeply declining value of marginal reductions with a falling opportunity cost of abatement. Therefore, although toxics may have large absolute negative impacts on welfare, this argues for their early, but not necessarily aggressive, abatement.

And finally how does the income elasticity of the demand for environmental quality (ε) affects the onset and pace of regulation? We have already shown that the restriction of $\varepsilon > 1$ is not needed to generate sustainable growth. This parameter does however have a role to play in determining the timing of regulation. To illustrate its role consider the fast growth regime and let the gross marginal product of capital, A, rise. This necessarily raises g and if $\varepsilon > 1$, the Switching Locus shifts in. Abatement is hastened. When $\varepsilon < 1$, abatement is delayed and peak pollution levels shift right.[72] A similar set of results holds for increases in the productivity of abatement although there is an additional conflicting force. Therefore, in contrast to earlier work Brock and Taylor (2003) finds that the income elasticity of marginal damage has an important role to play in determining the income level at which abatement occurs and the resulting pollution level, but virtually no role in determining if the environment will improve nor its rate of improvement.

6 Conclusion and Suggestions for Future Research

The relationship between economic growth and the environment is not well understood: we have only limited understanding of the basic science involved – be it physical or economic – and we have very limited data. In this review we have tried to evaluate ongoing efforts, both theoretical and empirical, to understand this relationship. We started by introducing definitions for the scale, composition and technique effects of growth on pollution, and then constructed three simple theoretical models to highlight the role each can play in generating sustainable growth. Throughout we have tried to link these models to the existing literature and in a very rudimentary way evaluated their predictions using data on pollution emissions, abatement costs and resource prices.

This is a research topic on the periphery of growth theory proper. It placement reflects the lack of a core model to work with and the paucity of data for empirical analyses. This is unfortunate because an

of sewage into the Juan de Fuca Strait. Scientific studies have long argued that since the sewage is pumped through long outfalls into cold, deep, fast moving water there is no need for treatment. The CRD has always used these studies to delay building a treatment facility. Current plans are for secondary treatment to begin in 2020, but until then over 40 square kilometers of shoreline remains closed to shell fishing. Background information can be found at the Sierra Legal Defense fund site http://www.sierralegal.org/m_archive/1998-9/bk99_02_04.htm

[72] Our use of the terms delayed or hastened does not refer to calendar time, but rather to whether actions occur at a higher or lower income level.

understanding of the relationship between economic growth and the environment may be key to long run prosperity; it is certainly of interest to developing country governments searching for a balance between material growth and environmental protection, and it is also of great interest in the developed world given current debates over global warming, its costs, and the costs of its amelioration.

Our review has revealed much heterogeneity in terms of approach and methods used in theoretical work. Some heterogeneity is to be expected, but too much dissipates effort. By examining the pollution creation and abatement process in some detail we hoped to direct future efforts more productively. We showed that standard assumptions such as CRS and concavity of the abatement production function lead to tractable formulations where pollution emissions appear in much the same way as other factors. By doing so we were able to show how we can evaluate the costs of environmental policy in a manner similar to that used to evaluate the drag of natural resources on growth. By making this connection precise we provided a bridge between the early resources and growth literature of the 1970s with the recent literature on pollution and income. We also hope to instill in others the need to provide micro foundations for assumptions over the amount of pollution emitted or abated in production, since we have repeatedly shown the importance of these assumptions for a model's ability to generate sustainable growth.

Our theoretical review contains three main messages. The first comes directly from our Green Solow model where we showed how the typical convergence properties of the neoclassical model together with a standard natural regeneration function yield an Environmental Kuznets Curve. This suggests that efforts to explain the EKC via complicated processes of political economy, IRS, freer trade, and differential factor growth, etc. may be unnecessary. At the very least it points out that the interplay of natural and Solow growth dynamics certainly work towards this finding.

Our second message concerns drag. We have shown throughout that efforts to limit pollution and raise environmental quality create a drag on growth rates. This finding was stronger in some cases since rapid population growth could eliminate the possibility of sustainable growth entirely. The drag calculations we provided are for illustration and not meant to substitute for more serious enquiry that must include empirical estimation of key parameters. Nevertheless these calculations are helpful in focusing our efforts on key parameters (the share of emissions in final good production or the rate of change in pollution abatement costs), and demonstrate how difficult it is to generate sustainable growth in a country with significant population growth. The calculations also offer a quick litmus test; if a specification suggests environmental policy reduces growth by 40%, something is surely amiss. It is hoped that drag calculations of the type we have conducted become a more standard feature in the literature.

Finally, we have shown how different assumptions on abatement can produce very different results for sustainability (recall the contradictory results of Smulders and Stokey in the AK model). To a certain extent progress in this literature has been slowed because researchers have too many degrees of freedom in choosing their specification. Some restrictions are imposed by the requirement of a balanced growth path, but this still leaves much leeway to the researcher. We have adopted a consistent specification of pollution creation and abatement based on the common, if not innocuous, assumptions of constant returns, concavity and pollution being a joint product of output. Within these confines we have then argued that technological progress in abatement, distinct from that in final goods, is key to generating sustainable growth at reasonable costs. By identifying this as a key requirement we hope to direct future research efforts towards a theory of induced innovation where both relative prices and pollution regulations determine the pace and direction of improvements in abatement technology.

Our review of empirical work shows that the existing literature has made relatively few contributions to our understanding. The Environmental Kuznet's curve stands out as a key empirical regularity, but continued progress in this area can only come with a more serious consideration of other related data. One contribution of this review has been to show that many of the theoretical models capable of generating an EKC also contain predictions in other directions that are worthy of examination. The simple Green Solow model had strong predictions for the emission intensity of GDP; the Stokey Alternative contained sharp predictions about the time profile of abatement costs; the Source-and-Sink model contained links between energy prices, energy use, and pollution levels; and finally, the Kindergarten model produced a cross-country catch-up hypothesis as well as yielding several within-country but across-pollutant predictions. Further

progress in our understanding can only come from a tighter connection between theory and data.

This review has been limited by its focus. It has been a review of work linking industrial pollution and growth with only small asides to consider natural resource use. In many cases the formal structure of the models resembled those in the renewable resource literature, but we did not provide a review of findings there. As such we have sidestepped the rather thorny issues of property rights protection and the efficiency of environmental policies. We have done so not because we believe that these issues do not merit attention, but rather because adding a useful discussion of these topics would make this review unwieldy. It should be emphasized however that a common feature of the resources we examined was their well-defined property rights. This is true for air quality when the quality is determined by local pollution; and it is true of oil and other energy resources.

There are however an important class of resources where property rights enforcement is lax or where no property rights exist at all. Property rights problems arise in three main areas. These are: local and transnational fisheries; the global atmospheric commons; and lastly, the forest stocks in many developing countries. It is somewhat ironic that these renewable resources are under far more threat than the so-called exhaustible resources such as oil, gas or minerals. The reason for this is inescapable: the diffuse nature of many of these resources has led to a lack of property rights and very little management. Therefore, while our focus on industrial pollution is perhaps defensible in that it determines the air quality and health prospects for hundreds of millions of people across the globe, we should not forget other vexing problems arising from the lack of property rights. And while our data and the existing empirical results suggest that many local pollution problems are well in hand or respond well to increases in incomes brought about by growth, global pollution problems, such as global warming, appear to be far more difficult to solve.[73] Therefore, it may be that the real threat to continued growth arises not from the relatively small drag introduced by existing environmental policies, but from the absence of new policies to stem more serious global problems.

[73] See Schmalensee, Stoker, and Judson (1998) and Holtz-Eakin and Selden (1995).

References

[1] Aghion, P. and P. Howitt (1998), Endogenous Growth Theory, MIT Press, Cambridge MA.

[2] Andreoni, J. and A.Levinson (2001), "The Simple Analytics of the Environmental Kuznets Curve," Journal of Public Economics, May , 80(2): 269-286.

[3] Antweiler, A., B. Copeland and M. S. Taylor (2001), "Is Free Trade Good for the Environment," American Economic Review 94, 1, September: 877-908.

[4] Arrow, K. (1962), "The Economic Implications of Learning by Doing" The Review of Economic Studies, Vol. 29, No. 3. June: 155-173.

[5] Arrow, K., P. Dasgupta, L. Goulder, G. Daily, P. Ehrlich, G. Heal, S. Levin, K. Maler, S. Schneider, D. Starrett, B. Walker (2003), "Are we consuming too much?", at P. Dasgupta's website, Cambridge.

[6] Barbier, E.(1997), "Environmental Kuznets Curve Special Issue: Introduction", Environment and Development Economics 2: 369-381.

[7] Barnett, H.J. and C. Morse (1963), Scarcity and Growth: The Economics of Natural Resource Availability, John Hopkins Press, Baltimore.

[8] Barret, S. and K. Graddy (2000), "Freedom, Growth, and the Environment", Environment and Development Economics. Vol. 5: 433-456

[9] Berndt, E.R.(1990), "Energy Use, Technical Progress and Productivity Growth: A Survey of Economic Issues", The Journal of Productivity Analysis, 2: 67-83 .

[10] Bovenberg, A.L., and S. Smulders (1995), "Environmental quality and pollution augmenting technological change in a two sector endogenous growth model", Journal of Public Economics, 57: 369-391.

[11] Brander, J.A., and M. S. Taylor (1997), "The Simple Economics of Easter Island: A Ricardo-Malthus model of renewable resource use", American Economic Review 88, No. 1: 119-138.

[12] Brock, W.A. (1977), "A Polluted Golden Age", in V.L. Smith, ed. Economics of natural and environmental resources (Gordon and Breach Science Publishers, London).

[13] Brock, W.A. and D. Starrett (2003), "Managing systems with nonconvex positive feedback,", Environmental and Resource Economics, 26, Iss. 4, Dec.: 575-624

[14] Clark, C. (1990). Mathematical Bioeconomics, Wiley.

[15] Brock, W.A. and M.S. Taylor (2004),. "The Green Solow model". NBER Working Paper No. w10557. June.

[16] Brock, W.A. and M.S. Taylor (2003), "The Kindergarten Rule for Sustainable Growth". NBER Working Paper No. w9597. April .

[17] Cole, M.A., A.J. Rayner and J.M. Bates (1997), "The environmental Kuznets curve: an empirical analysis", Environmental and Development Economics, 2, 401-416.

[18] Copeland, B.R., and Taylor M.S. (1994), "North-South Trade and the Environment", Quarterly Journal of Economics, 109: 755-787.

[19] Copeland, B.R., and Taylor M.S. (2003), Trade and the Environment: Theory and Evidence, Princeton University Press, Princeton, NJ.

[20] Copeland, B.R., and Taylor M.S. (2004), "Trade, Growth and the Environment", Journal of Economic Literature, forthcoming.

[21] Dasgupta, P, and G. Heal (1974), "The Optimal Depletion of Exhaustible Resources", The Review of Economics Studies, Vol. 41, Symposium Issue on the Economics of Exhaustible Resources: 3-28.

[22] Dasgupta, P, and G. Heal (1979), "Economic Theory and Exhaustible Resources, Cambridge Economic Handbooks, Cambridge University Press.

[23] Dasgupta, P. and K. Maler (2003), "The economics of non-convex ecosystems: Introduction," Environmental and Resource Economics, 26, Iss. 4, Dec.: 499-602 .

[24] Dechert, W.D. e.d. (2001), Growth Theory, nonlinear dynamics and Economic modeling, Edward Elgar: Cheltenham.

[25] Durlauf, S.N. and D.T. Quah (1999), "The New Empirics of Economic Growth", in the Handbook of Macroeconomics 1, chapter 4, J.B. Taylor and M. Woodford, Eds. Elsevier Science B.V.:235-308.

[26] EPA (2004), Strategic Plan, Benefits the Costs of EPA's Activities available at http://www.epa.gov/ocfo/plan/plan.htm

[27] EPA (1990), Environmental Investments: The Cost of a Clean Environment.

[28] Forster, B.A. (1973), "Optimal Capital Accumulation in a Polluted Environment" Southern Economic Journal, 39: 544-547

[29] Gale L. and Mendez J.A. (1998), "A note on the relationship between, trade, growth, and the environment". International Review of Economics and Finance 7(1): 53-61

[30] Goulder, L.H. and Schneider, S.H. (1999), "Induced Technological change and the Attractiveness of CO2 abatement policy", Resource and Energy Economics, August, 21 (3-4): 211-253.

[31] Grossman G.M, and A. B. Krueger (1993), "Environmental Impacts of a North American Free Trade Agreement," in The US-Mexico Free Trade Agreement, P. Garber, ed. Cambridge, MA: MIT Press.

[32] Grossman G.M, and A. B. Krueger (1995), "Economic Growth and the Environment," Quarterly Journal of Economics: 353-377.

[33] Grossman G.M. and Helpman E. (1991), Innovation and Growth in the Global Economy. MIT press.

[34] Harbaugh, W. Arik Levinson, and David Molloy Wilson (2002), "Reexamining the Empirical Evidence for an Environmental Kuznets Curve". The Review of Economics and Statistics.84(3) August: 541-551.

[35] Hartwick, J.M. (1977), "Intergenerational Equity and the Investing of Rents from Exhaustible Resources, American Economic Review, 66, December.

[36] Hilton, H., and A. Levinson (1998), "Factoring the Environmental Kuznets Curve: Evidence from Automotive Lead Emissions," Journal of Environmental Economics and Management, 35: 126-141.

[37] Holtz-Eakin, D. and T. Selden (1995), "Stoking the fires? CO2 emissions and economic growth," Journal of Public Economics, 57: 85-101.

[38] Hotelling, H. (1931) "The Economics of Exhaustible Resources", Journal of Political Economy, 39 April: 137-175.

[39] Jaffe, A.B., K. Palmer (1997), "Environmental Regulation and Innovation: A panel data study", Review of Economics and Statistics, Nov.79(4): 610-619.

[40] Jaffe, A.B., S.R. Peterson, and P.R. Portney (1995), "Environmental Regulation and the Competitiveness of U.S. Manufacturing: What does the Evidence tell us?, Journal of Economic Literature, Vol. XXXIII, March: 132-163.

[41] Jaffe, A. B. and R.G. Newell, and R.N. Stavins (forthcoming), "Technological change and the Environment" in the Handbook of Environmental Economics, Karl-Goran Maler and Jeffrey Vincent eds.. North-Holland/Elsevier Science, eds.

[42] John, A. and R. Pecchenino (1994), "An overlapping generations model of growth and the environment," The Economic Journal, 104: 1393-1410.

[43] Jones C.I. (2001), Introduction to Economic Growth. Second Edition. W.W. Norton & Company.

[44] Jones, L.E. and R.E. Manuelli (2001), "Endogenous Policy Choice: The Case of Pollution and Growth," , Review of Economic Dynamics, 4, Issue 2, April: 245-517.

[45] Jorgenson, D.W. and P.W. Wilcoxen (1990), "Environmental Regulation and U.S. economic growth", Rand Journal of Economics, 21, summer: 314-340.

[46] Kaufman, Robert F.(2004), "The Mechanisms for Autonomous Energy Efficiency Increases: A Cointegration Analysis of US Energy/GDP Ratio", The Energy Journal, 25, No. 1.

[47] Keeler, E., M. Spence, and R. Zeckhauser (1972), "The Optimal Control of Pollution", Journal of Economic Theory, 4: 19-34.

[48] Levinson, A., Taylor M. S. (2003), "Unmasking the Pollution Haven Effect", paper presented at the NBER Environmental Economics Meetings, Boston, Summer. Available at http://www.ssc.wisc.edu/~staylor/

[49] Lopez, R. (1994), "The environment as a factor of production: the effects of economic growth and trade liberalization", Journal of Enviornmental Economics and Management 27: 163-184.

[50] Lucas, R.E. Jr. (1998), "On the Mechanics of Economic Development," Journal of Monetary Economics, 22: 3-42.

[51] Meadows, D.H., Meadows, D.L., Randers, J., Behrens, W.W. (1972), The limits to Growth, Universe Books, New York.

[52] Meadows, d.H., Meadows, D.L., Randers, J. (1991), Beyond the Limits, Earthscan Publications, London.

[53] McConnell, K.E. (1997), "Income and the demand for environmental quality", Environment and Development Economics 2: 383-399.

[54] Newell, R.G., A.B. Jaffe, and R.N. Stavins (1999), "The Induced Innovation hypothesis and energy saving technological change", Quarterly Journal of Economics, August: 941-975.

[55] Nordhaus, W. (1992), "Lethal Model 2: The Limits to Growth Revisited" Brookings Papers on Economic Activity, Volume , No. 2: 1-59.

[56] Pezzey, J.C.V., and Withagen, C. (1998), "The Rise, Fall and Sustainability of Capital-Resource Economies", Scandinavian Journal of Economics, 100, (2): 513-527.

[57] Popp, D. (2002), "Induced Innovation and Energy Prices", American Economic Review, Vol. 92, No. 1, March: 160-180.

[58] Rodriguez, F., and J.D. Sachs (1999), "Why Do Resource Abundant Economies Grow more slowly", Journal of Economic Growth, 4, September: 277-303.

[59] Robinson, J.A. and T.N. Srinivasan (1997), "Long-term consequences of population growth: technological change, natural resources, and the environment", Handbook of Population and Family Economics, eds. M.R. Rosenzweig and O. Stark, Elsevier Science.

[60] Romer, P. (1986), "Increasing Returns and Long Run Growth", Journal of Political Economy, 94, October: 1002-37.

[61] Rose, A. and C.Y. Chen (1991), "Sources of change in energy use in the U.S. economy, 1972-1982: a structural decomposition analysis", Resources and Energy 13: 1-21.

[62] Schmalensee, R., Stoker T.M. and Judson R.A. (1998), "World Carbon Dioxide Emissions: 1950 - 2050", The Review of Economics and Statistics, 80, Issue 1, February: 15-27

[63] Scheffer, M., S.R.Carpenter (2003), "Catastrophic regime shifts in ecosystems: linking theory to observation", Trends in Ecology and Evolution, Vol. 18, No. 12, December: 648-656.

[64] Scholz, C.M. and G. Ziemes (1999), "Exhaustible Resources, Monopolistic Competition, and Endogenous Growth", Environmental and Resource Economics, 13: 169-185.

[65] Selden T. and D. Song (1994), "Environmental quality and development: Is there a Kuznets curve for air pollution emissions?" Journal of Environmental Economics and Management 27: 147-162.

[66] Shafik N. and S. Bandyopadhyay (1994), "Economic Growth and Environmental Quality: time series and cross country evidence," Washington D.C: The World Bank.

[67] Slade, M.E. (1987), "Trends in Natural Resource commodity prices: an analysis in the time domain", Journal of Environmental Economics and Management, 9: 122-137.

[68] Smulders, J. (1994), Growth, Market Structure, and the Environment: Essays on the Theory of Endogenous Economic Growth. Tilburg University.

[69] Smulders, S. (1999), "Endogenous Growth Theory and the Environment", in J.C.J.M. van den Berg (eds.) Handbook of Enviornmental and Resource Economics, Cheltenham: Edward Elgar: 610-621.

[70] Smulders, S. and R. Gradus.(1996), "Pollution abatement and long-term growth", European Journal of Political Economy, 12: 505-532.

[71] Solow, R.M. (1974), "Intergenerational Equity and Exhaustible Resources", Review of Economic Studies (Symposium): 29-46.

[72] Solow, R. (1973),"Is the end of the world at hand", Challenge, March-April: 39-50.

[73] Solow, R.M., (1993), "An almost practical step toward sustainability", Resources Policy 19(3); 162-172.

[74] Stiglitz, J. (1974), "Growth with Exhaustible Natural Resources: Efficient and Optimal Growth Paths" The Review of Economic Studies, Vol. 41, Symposium on the Economics of Exhaustible Resources: 123-137.

[75] Stern, David, I. (2003), "The rise and fall of the environmental kuznets curve," Department of Economics, Rensselaer Polytechnic Institute, http:www.rpi.edu/dept/economics/.

[76] Stokey, N.(1998), "Are there limits to Growth," International Economic Review, 39(1): 1-31

[77] Sue Wing, Ian, and R.S. Eckhaus (2003), "The Energy Intensity of US Production: Sources of Long-Run Change, paper presented at the 5th USAEE/IAEE session of the Allied Social Science Association meeting, Washington, D.C. Jan. 4.

[78] Tahvonen, O., and J. Kuuluvainen (1991), "Optimal Growth with renewable resources and pollution", European Economic Review, 35: 650-661.

[79] Valente, S. (2002), "Renewable Resources and Sustainable Development", mimeo, Universita di roma.

[80] Vincent, J.R. (1997), "Testing for environmental Kuznets curves within a developing country", Special Issue on Environmental Kuznets Curves, Environment and Development Economics 2 (4): 417-431.

[81] Weitzman, M. L.(1999), "Pricing the Limits to Growth from Mineral Depletion", Quarterly Journal of Economics, May: 691-706.

FIGURES

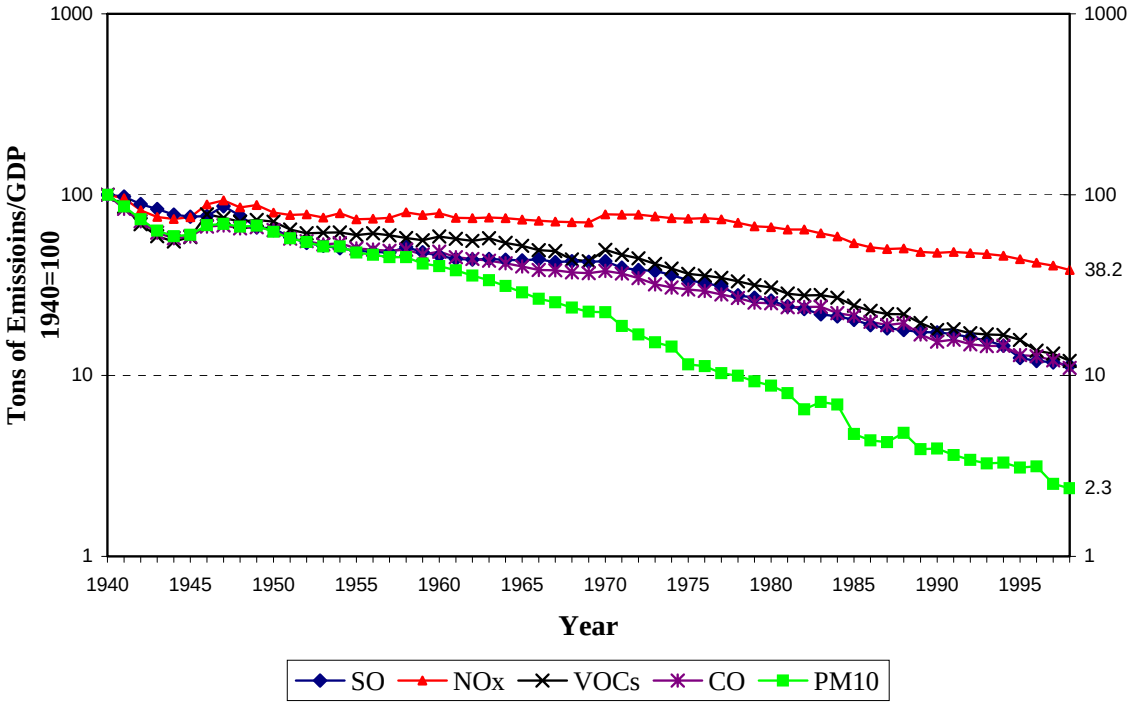


Figure 1: Emission intensities, 1940-1998. Tons of emissions/real GDP.

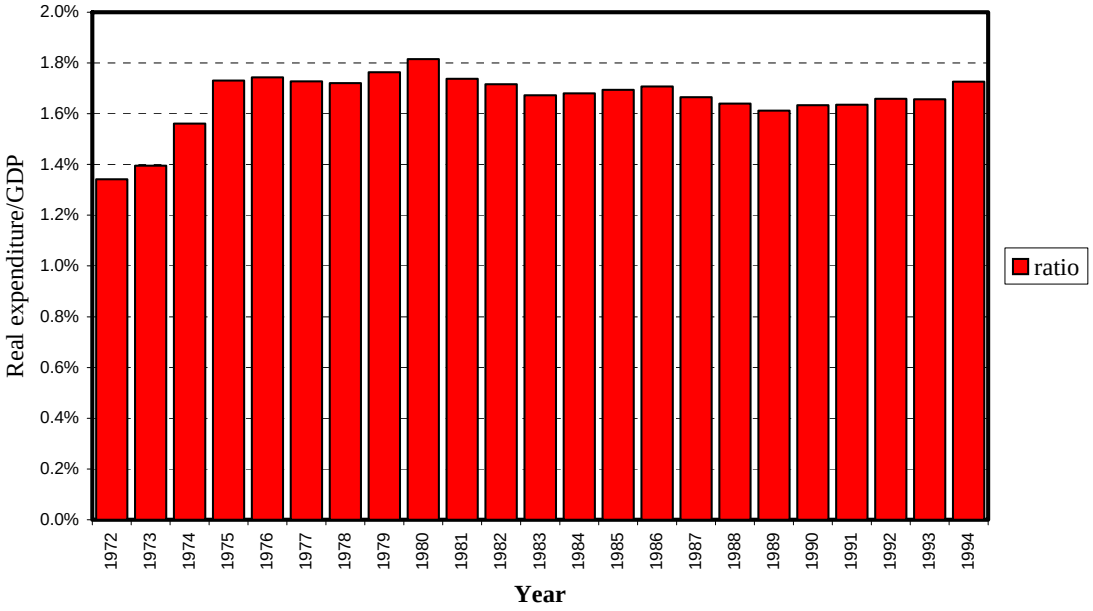


Figure 2: Pollution abatement costs, 1972-1994. PACE/GDP.

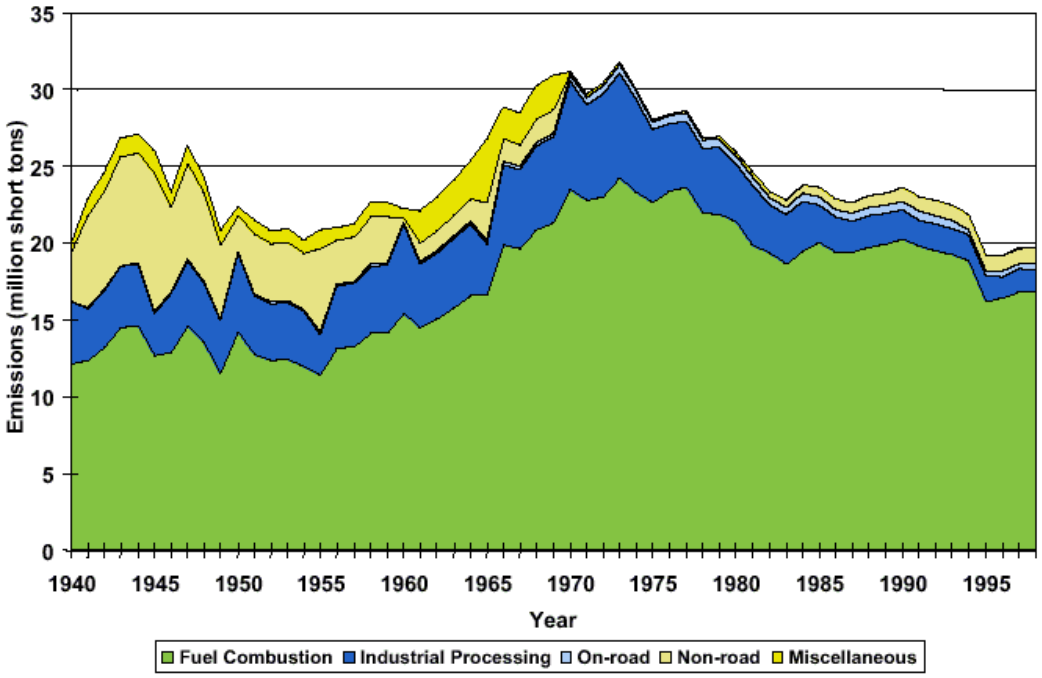


Figure 3: Sulfur dioxide emissions, 1940-1998.

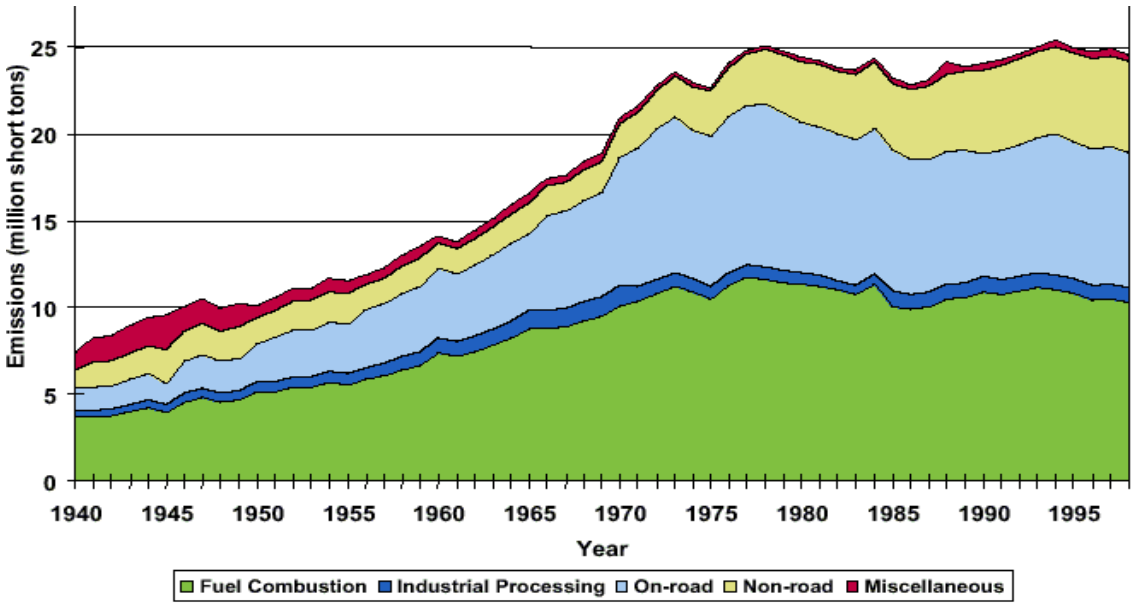


Figure 4: Nitrogen oxide emissions, 1940-1998.

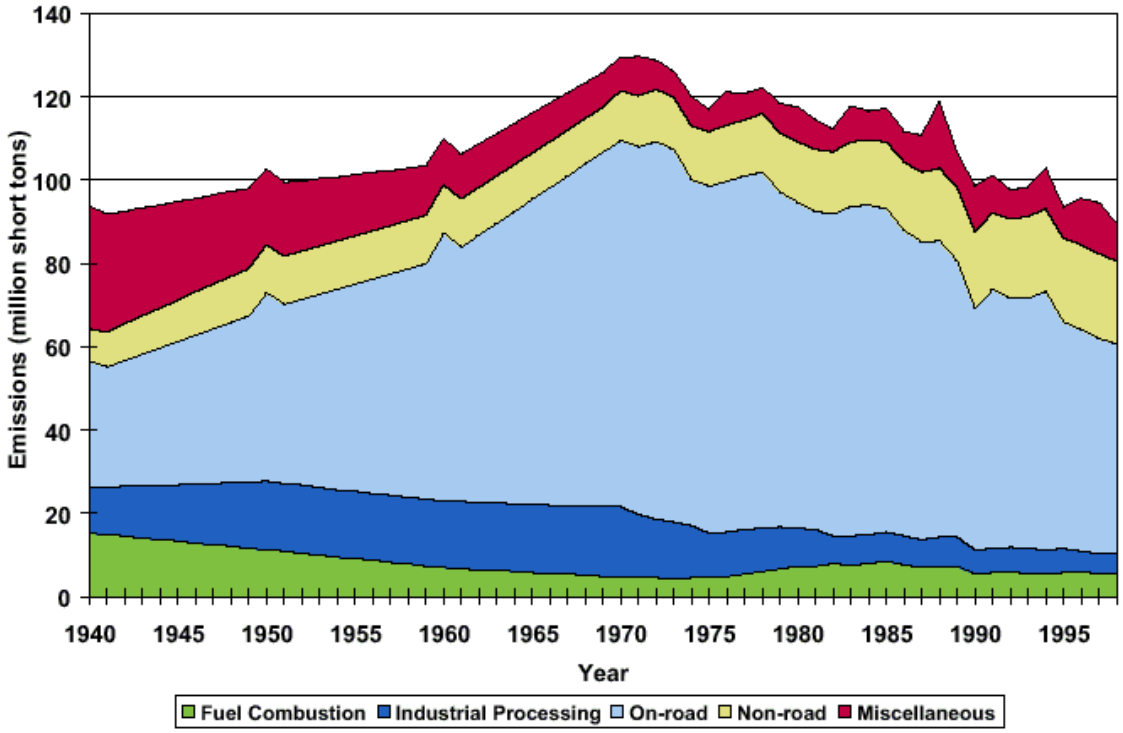


Figure 5: Carbon monoxide emissions, 1940-1998.

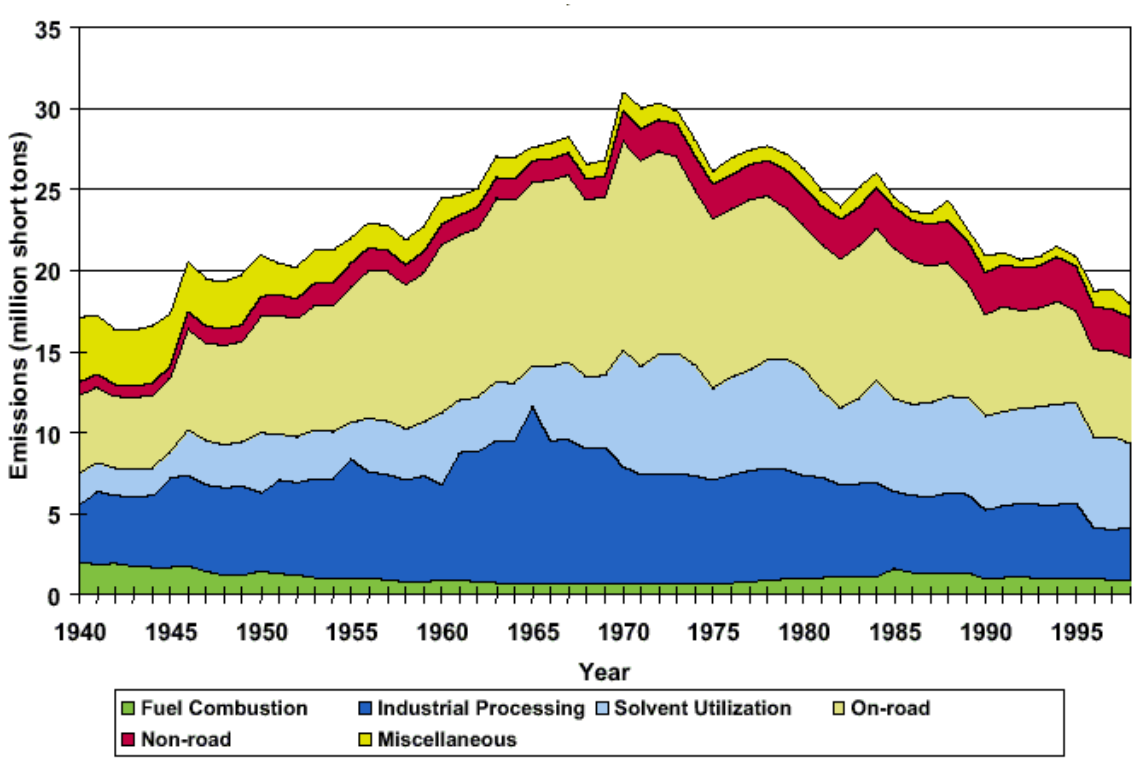


Figure 6: Volatile organic compounds, 1940-1998

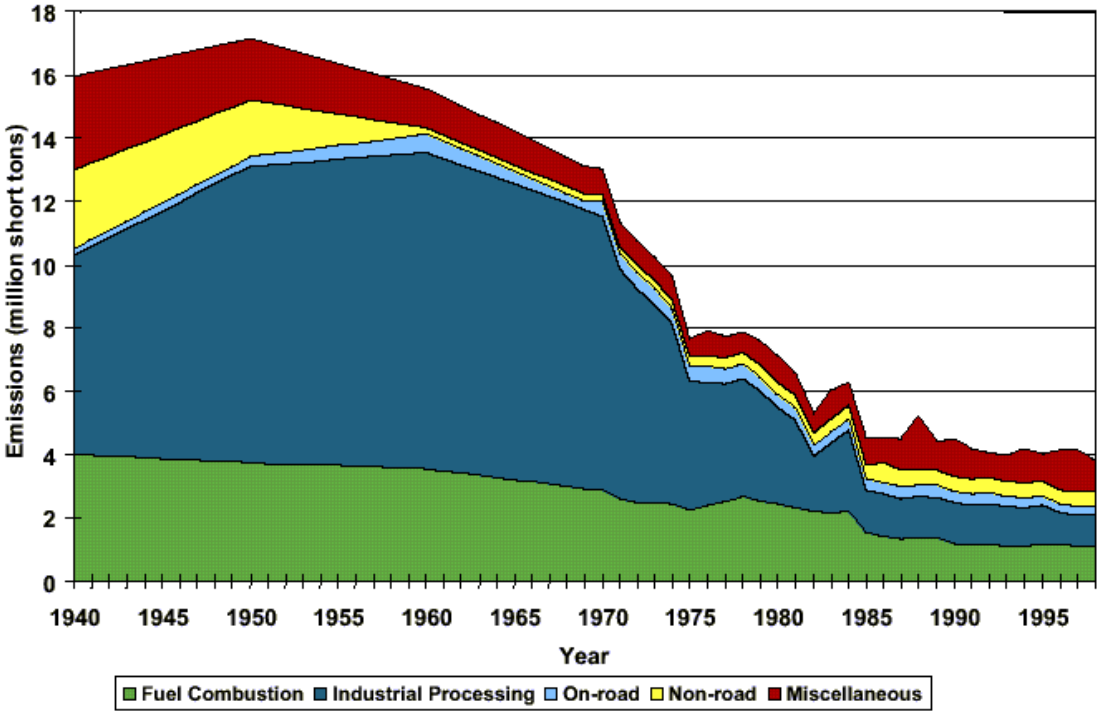


Figure 7: Particular matter PM10, 1940-1998.

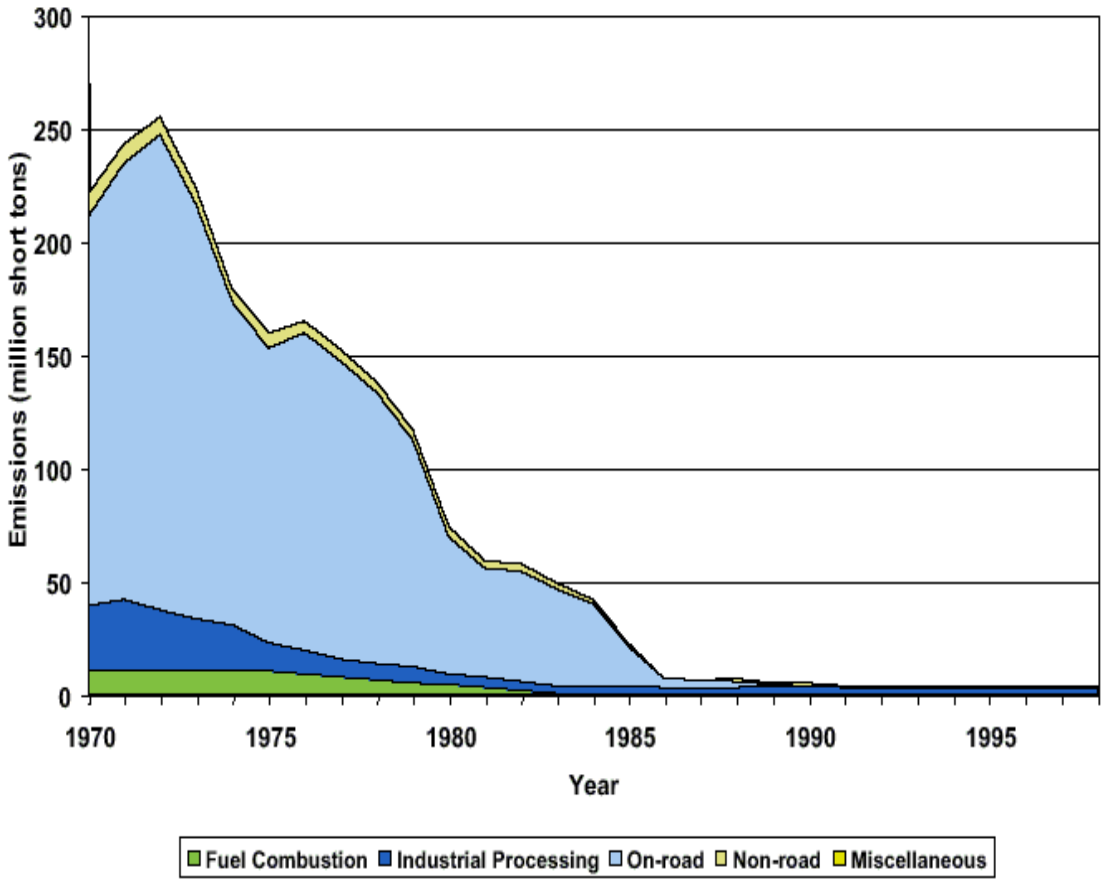


Figure 8: Lead emissions 1970-1998.

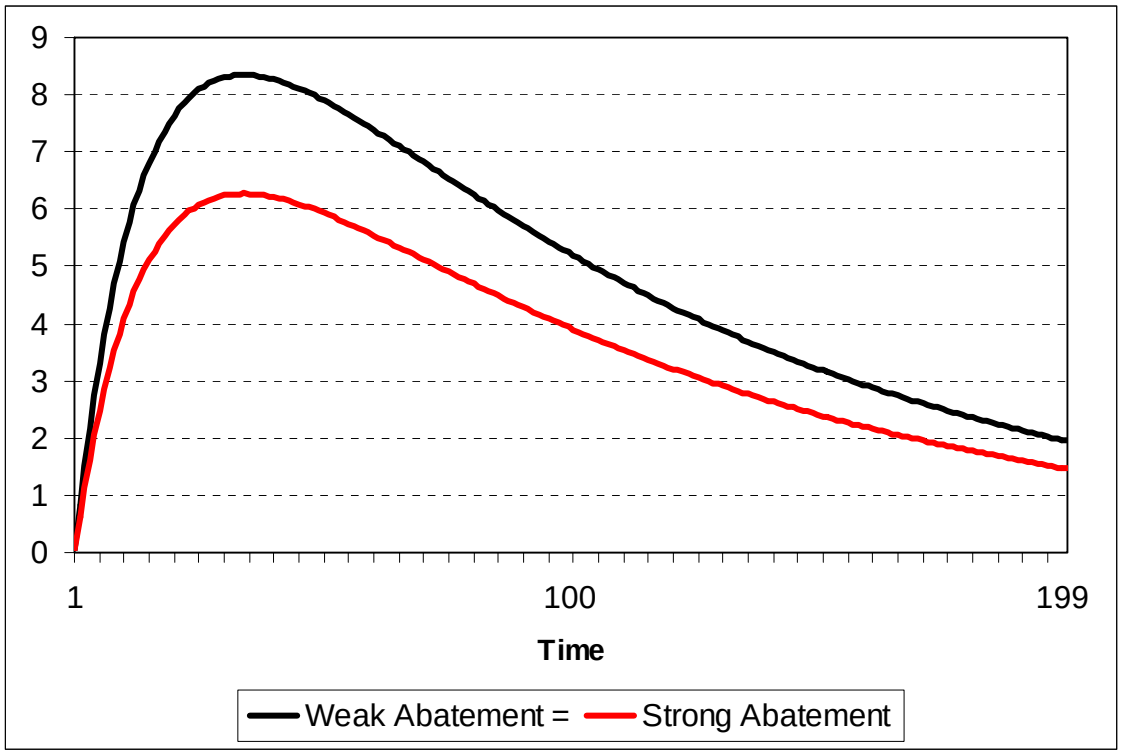


Figure 9: The green solow benchmark.

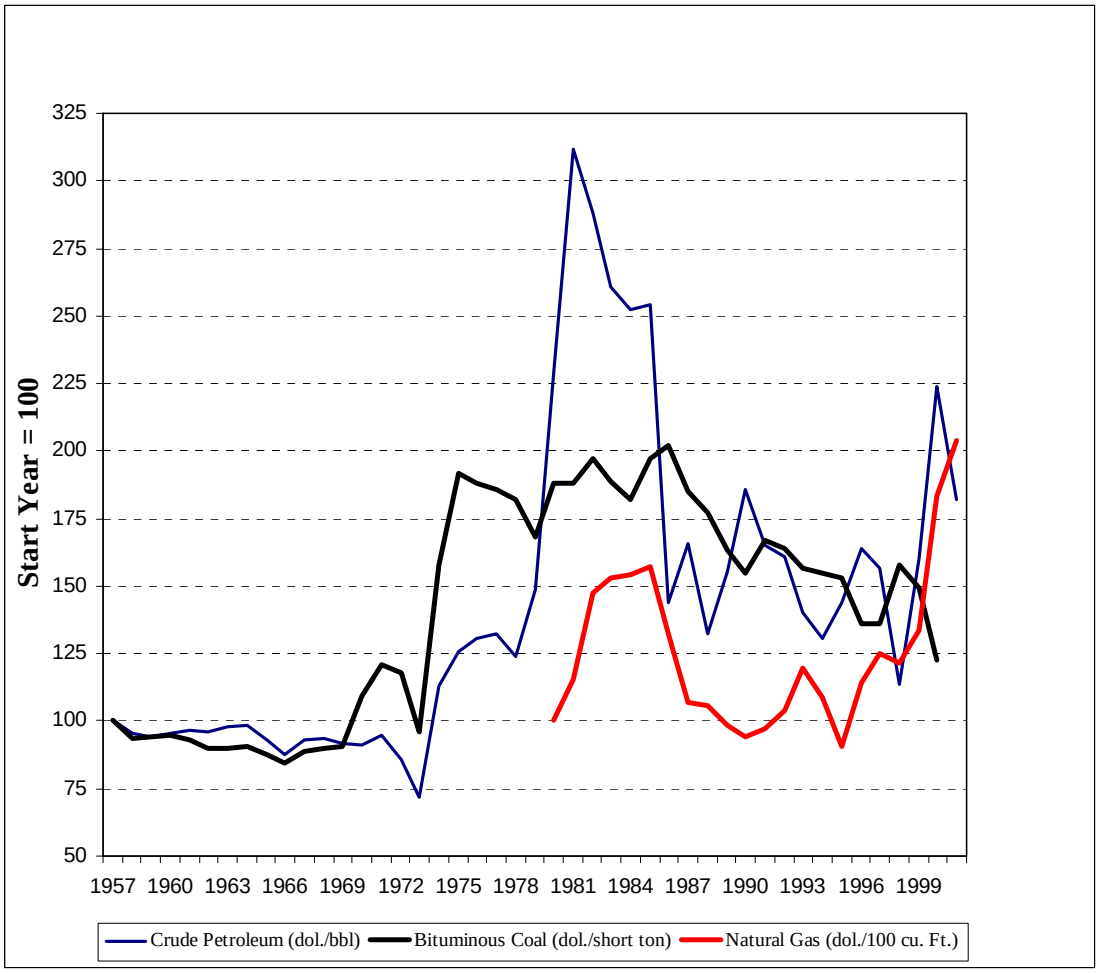


Figure 10: Real energy prices.

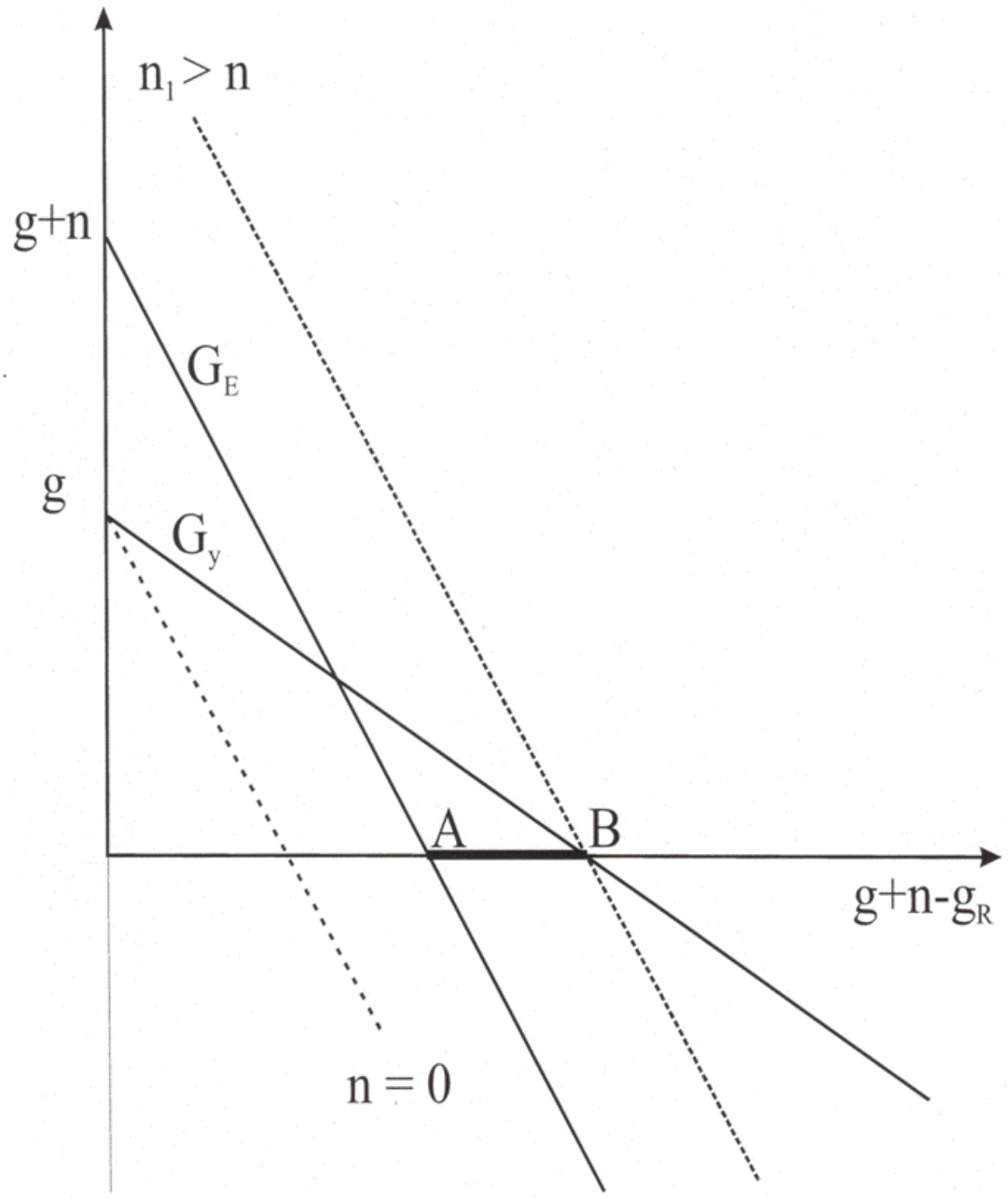


Figure 11: Feasibility: resource drag per capita growth.

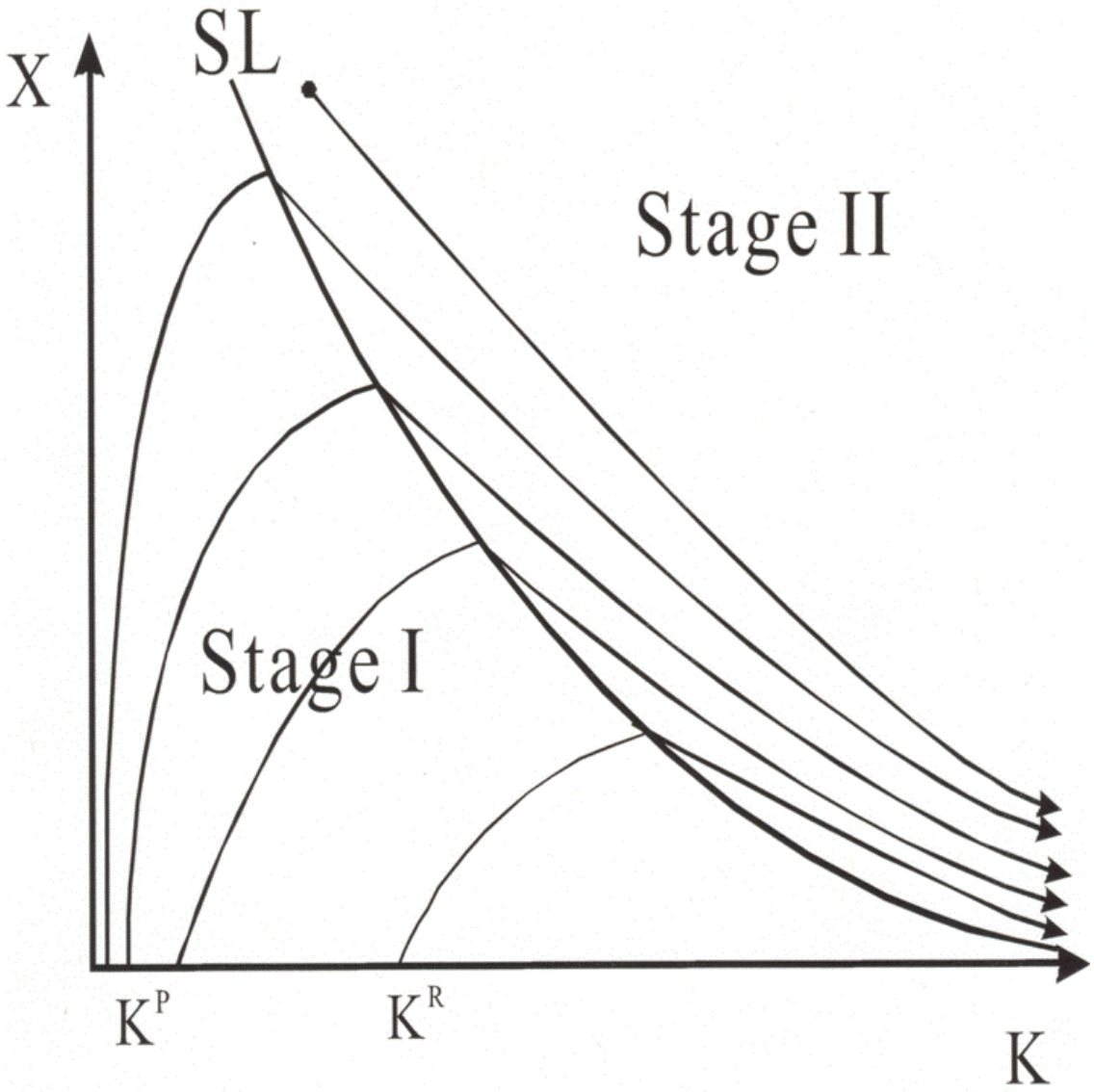


Figure 12: Transition paths.