SYSTEMS ANALYSIS

Does the system work? Is it worth the cost? Can and should it be implemented elsewhere? It is the reputed purpose of evaluation to provide answers to these and related questions. The need for conducting evaluations becomes more critical as systems or programs become more complex and more costly and, concomitantly, as the tax base or resources for their funding remain fixed or decrease. Unfortunately, system or program evaluation has not lived up to expectations (1). The field of evaluation is littered with efforts that do not adequately address the important issues or objectives, that do not employ valid controls for comparison purposes, that rely on inadequate measures or include expensive collections of data on measures that are in fact never used in the evaluation, that rely on inappropriate measurement methods, or that employ inadequate analytic techniques. Most, if not all, of the above-cited problems could be mitigated by developing, at the beginning of an evaluation effort, a valid and comprehensive evaluation design.

Although there is no stock evaluation design that can be taken off the shelf and implemented without revision, there should be an approach or process by which such designs can be developed. Indeed, Tien (2) outlines a systems approach—that is at once purposeful and systematic—for developing valid and comprehensive evaluation designs. The approach was first proposed by Tien (3) and has since been successfully employed in a number of evaluation efforts [see, e.g., Colton et al. (4), Tien and Cahn (5), and Tien and Rich (6)]. The approach is outlined in the next section, followed first by an illustration of the importance of evaluation modeling, and then by an observation that what is also needed is a continuous layered approach to the monitoring, diagnosis, and improvement of systems that could complement the broader system evaluations that, by necessity, are undertaken on an intermittent and as-needed basis.

EVALUATION APPROACH

The evaluation approach is based on a dynamic roll-back framework that consists of three steps leading up to a valid and comprehensive evaluation design. The roll-back aspect of the framework is reflected in the ordered sequence of steps. The sequence rolls back in time from (1) a projected look at

the range of program characteristics (i.e., from its rationale through its operation and anticipated findings); to (2) a prospective consideration of the threats (i.e., programs and pitfalls) to the validity of the final evaluation; and to (3) a more immediate identification of the evaluation design elements. The logic of the sequence of steps should be noted; that is, the anticipated *program characteristics* identify the possible *threats to validity,* which in turn point to the *evaluation design elements* that are necessary to mitigate, if not to eliminate, these threats. The three-step sequence can also be stated in terms of two sets of links that relate, respectively, an anticipated set of program characteristics *to* an intermediate set of threats to validity *to* a final set of design elements. Although some of the links between program characteristics and threats to validity are obvious (e.g., a concurrent program may cause an extraneous event threat to internal validity), an exhaustive listing of such links—for purposes of, say, a handbook—will require a significant amount of analysis of past and ongoing evaluations. Similarly, the second set of links between threats to validity and design elements will also require a significant amount of analysis. Both sets of links are briefly considered herein.

The "dynamic" aspect of the framework refers to its nonstationary character; that is, the components of the framework must be updated constantly, throughout the entire development and implementation phases of the evaluation design. In this manner, the design elements can be refined, if necessary, to account for any new threats to validity that may be caused by previously unidentified program characteristics.

In sum, the dynamic roll-back framework is systems oriented; it represents a purposeful and systematic process by which valid and comprehensive evaluation designs can be developed. Each of the three steps in the design framework is elaborated on in the next three subsections, respectively.

Program Characteristics

In general, the characteristics of a program can be determined by seeking responses to the following questions: What is the program rationale? Who has program responsibility? What is the nature of program funding? What is the content of the program plan? What are the program constraints? What is the nature of program implementation? What is the nature of program operation? Are there any other concurrent programs? What are the anticipated evaluation findings?

Again, it should be noted that the purpose of understanding the program characteristics is to identify the resultant problems or pitfalls that may arise to threaten the validity of the final evaluation. The possible links between program characteristics and threats to validity are considered in the next subsection, following a definition of the threats to validity.

Threats to Validity

After more than three decades, the classic monograph by Campbell and Stanley (7) is still the basis for much of the ongoing discussion of threats to validity. However, their original 12 threats have been expanded by Tien (3) to include 8 additional threats. The 20 threats to validity can be grouped into the following five categories.

1. *Internal validity* refers to the extent that the statistical association of an intervention and measured impact can reasonably be considered a causal relationship. This category includes the following 9 threats: (1) extraneous events, (2) temporal maturation, (3) design instability, (4) pretest experience, (5) instrumentation changes, (6) regression artifacts, (7) differential selection, (8) differential loss, and (9) selection-related interaction.
2. *External validity* refers to the extent that the causal relationship can be generalized to different populations, settings and times. This category includes the following 4 threats: (10) pretest intervention interaction, (11) selection-intervention interaction, (12) test-setting sensitivity, and (13) multiple-intervention interference.
3. *Construct validity* refers to the extent that the causal relationship can be generalized to different interventions, impact measures, and measurements. This category includes the following 2 threats: (14) intervention sensitivity, and (15) measures sensitivity.
4. *Statistical conclusion validity* refers to the extent that an intervention and a measured impact can be statistically associated: error could be either a false association (i.e., Type I error) or a false nonassociation (i.e., Type II error). This category includes the following 2 threats: (16) extraneous sources of error, and (17) intervention integrity.
5. *Conduct conclusion validity* refers to the extent that an intervention and its associated evaluation can be completely and successfully conducted. This category includes the following 3 threats: (18) design complexity, (19) political infeasibility, and (20) economic infeasibility.

Although the 20 threats to validity are, for the most part, self-explanatory, it is helpful to highlight three aspects. First, the threats to external and construct validities are threats to the *generalizability* of the observed impacts. Generalization involves the science of induction, which causes a number of problems that are, according to Campbell and Stanley (7, p. 17),

> painful because of a recurrent reluctance to accept Hume's truism that *induction or generalization is never fully justified logically.* Whereas the problems of *internal* validity are solvable within the limits of the logic of probability and statistics, the problems of external validity are not logically solvable in any near, conclusive way. Generalization always turns out to involve extrapolation into a realm not represented in one's sample. Such extrapolation is made by *assuming* one knows the relevant laws.

Although generalization is difficult to undertake, it is a fundamental aspect of social program evaluation. While the classical sciences (i.e., physics, chemistry, biology, etc.) emphasize *repeatability* in their experiments, the social sciences emphasize *representativeness* in their experiments, thus facilitating extrapolations or generalizations.

Second, it can be seen that the threats to validity identified above are overlapping in some areas and conflicting in other areas. For example, seasonal effects could be identified either as extraneous events or a result of temporal maturation. Additionally, factors that mitigate threats to conduct conclusion validity would most likely be in conflict with those that mitigate the other threats to validity. It is, however, *essential* that the threats to conduct conclusion validity be borne in mind

when developing an evaluation design; the field of evaluation is littered with studies that were not concluded because of the design's complexity or because of the political and economic infeasibilities that were initially overlooked.

Third, the threats to validity can be regarded as plausible rival hypotheses or explanations of the observed impacts of a program. That is, the assumed causal relationships (i.e., test hypotheses) may be threatened by these rival explanations. Sometimes the threats may detract from the program's observed impacts. The key objective of an evaluation design is then to minimize the threats to validity, while at the same time to suggest the causal relationships. The specific evaluation design elements are considered next.

Evaluation Design Elements

Tien (3) has found it systematically convenient to describe a program evaluation design in terms of five components or sets of design elements, including test hypotheses, selection scheme, measures framework, measurement methods, and analytic techniques.

Test Hypotheses. The test hypotheses component is meant to include the range of issues leading up to the establishment of test hypotheses. In practice, and as indicated in the dynamic roll-back framework, the test hypotheses should be identified only after the program characteristics and threats to validity have been ascertained.

The test hypotheses are related to the rationale or objectives of the program and are defined by statements that hypothesize the causal relationships between dependent and independent measures, and it is a purpose of program evaluation to assess or test the validity of these statements. To be tested, a hypothesis should (1) be expressed in terms of quantifiable measures, (2) reflect a specific relationship that is discernible from all other relations, and (3) be amenable to the application of an available and pertinent analytic technique. Thus, for example, in a regression analysis, the test hypothesis takes the form of an equation between a dependent measure and a linear combination of independent measures, while in a before-after analysis with a chi-square test, a simple test hypothesis, usually relating two measures, is used.

In the case of a complex hypothesis, it may be necessary to break it down into a series of simpler hypotheses that could each be adequately tested. In this manner, a measure that is the dependent measure in one test could be the independent measure in another test. In general, input measures tend to be independent measures, process measures tend to be both independent and dependent measures, while impact measures tend to be dependent measures.

Another difficulty arises in the testing process. Analytic techniques exist for testing the *correlation* of measures, but correlation does *not* necessarily imply *causation.* However, inasmuch as causation implies correlation, it is possible to use the relatively inexpensive correlational approach to weed out those hypotheses that do not survive the correlational test. Furthermore, in order to establish a causal interpretation of a simple or partial correlation, one must have a plausible *causal* hypothesis (i.e., test hypothesis) and at the *same* time *no* plausible *rival* hypotheses (i.e., threats to validity) that could explain the observed correlation. Thus, the fewer the number of plausible rival hypotheses, the greater is the likelihood that the test hypothesis is not *disconfirmed.* Alternatively, if a hypothesis is not disconfirmed or rejected after several independent tests, then a powerful argument can be made for its validity.

Finally, it should be stated that while the test hypotheses themselves cannot mitigate or control for threats to validity, poor definition of the test hypotheses can threaten statistical conclusion validity, since threats to validity represent plausible rival hypotheses.

Selection Scheme. The purpose of this component is to develop a scheme for the selection and identification of test groups and, if applicable, control groups, using appropriate sampling and randomization techniques. The selection process involves several related tasks, including the identification of a general sample of units from a well-designated universe; the assignment of these (perhaps matched) units to at least two groups; the identification of at least one of these groups to be the test group; and the determination of the time(s) that the intervention and, if applicable, the placebo are to be applied to the test and control groups, respectively. A more valid evaluation design can be achieved if random assignment is employed in carrying out each task. Thus, random assignment of units to test and control groups increases the comparability or equivalency of the two groups, at least before the program intervention.

There is a range of selection schemes or research designs, including *experimental* designs (e.g., pretest-posttest equivalent design, Solomon four-group equivalent design, posttest-only equivalent design, factorial designs), *quasi-experimental* designs (e.g., pretest-posttest nonequivalent design, posttest-only nonequivalent design, interrupted time-series nonequivalent design, regression-discontinuity design, ex post facto designs), and *nonexperimental* designs (e.g., case study, survey study, cohort study). In general, it can be stated that nonexperimental designs do not have a control group or time period, while experimental and quasi-experimental designs do have such controls even if it is just a before-after control. The difference between experimental and quasi-experimental designs is that the former set of designs have comparable or equivalent test and control groups (i.e., through randomization) while the latter set of designs do not.

Although it is always recommended that an experimental design be employed, there are a host of reasons that may prevent or confound the establishment—through random assignment—of equivalent test and control groups. One key reason is that randomization creates a focused inequity because some persons receive the (presumably desirable) program intervention while others do not. Whatever the reason, the inability to establish equivalent test and control groups should not preclude the conduct of an evaluation. Despite their inherent limitations, some quasi-experimental designs are adequate. In fact, some designs (e.g., regression-discontinuity designs) are explicitly nonrandom in their establishment of test and control groups. On the other hand, other quasi-experimental designs should be employed only if absolutely necessary and if great care is taken in their employment. Ex-post-facto designs belong in this category. Likewise, nonexperimental designs should only be employed if it is not possible to employ an experimental or quasi-experimental design. The longitudinal or cohort study approach, which is a nonexperimental design, is becoming increasingly popular.

In terms of selection scheme factors that could mitigate or control for the various threats to validity, it can be stated that randomization is the key factor. In particular, most, if not all, of the internal and external threats to validity can be mitigated by the experimental designs, which, in turn, can only be achieved through randomization. Thus, *random assignment* of units—especially *matched* units—to test and control groups can control for all the threats to internal validity except, perhaps, extraneous events, *random identification* of a group to be the test group and *random determination* of time(s) that the intervention is to be applied can control for selection-related interaction threats to internal validity, and *random sampling* can allow for generalization to the universe from which the sample is drawn.

Measures Framework. There are two parts to the measures framework component. First, it is necessary to specify the set of evaluation measures that is to be the focus of the particular evaluation. Second, a model reflecting the linkages among these measures must be constructed.

In terms of evaluation measures, Tien (3) has identified four sets of measures: input, process, outcome, and systemic measures. The input measures include program rationale (objectives, assumptions, hypotheses), program responsibility (principal participants, participant roles), program funding (funding level, sources, uses), program constraints (technological, political, institutional, environmental, legal, economic, methodological), and program plan (performance specifications, system design, implementation schedule). The process measures include program implementation (design verification, implementation cost), program operation (system performance, system maintenance, system security, system vulnerability, system reliability, operating cost), and concurrent programs (technological, physical, social). The outcome measures include attitudinal, behavioral, and other impact considerations. The systemic measures include organizational (intraorganizational, interorganizational), longitudinal (input, process, outcome), programmatic (derived performance measures, comparability, transferability, generalizability), and policy (implications, alternatives) considerations.

In general, the input and process measures serve to "explain" the resultant outcome measures. Input measures alone are of limited usefulness since they only indicate a program's potential, not actual, performance. On the other hand, the process measures do identify the program's performance but do not consider the impact of that performance. Finally, the outcome measures are the most meaningful observations since they reflect the ultimate results of the program. In practice and as might be expected, most of the available evaluations are fairly explicit about the input measures, less explicit about the process measures, and somewhat fragmentary about the outcome measures.

The fourth set of evaluation measures, the systemic measures, can also be regarded as impact measures but have been overlooked to a large extent in the evaluation literature. The systemic measures allow the program's impact to be viewed from at least four systemic perspectives. First, it is important to view the program in terms of the *organizational* context within which it functions. Thus, the program's impact on the immediate organization and on other organizations must be assessed. Second, the pertinent input, process, and outcome measures must be viewed over time, from a *longitudinal* perspective. That is, the impact of the program on a particular system must be assessed not only in comparison to an immediate "before" period but also in the context of a longer time horizon. Thus, it is important to look at a process measure like, for example, average response time over a five-to-ten-year period to ascertain a trend line, since a perceived impact of the program on the response time may be just a regression artifact. Third, in an overall *programmatic* context, the evaluator should (1) derive second-order, systems performance measures (e.g., benefit cost and productivity measures) based on the first-order input, process, and outcome measures; (2) compare the program results with findings of other similar programs; (3) assess the potential of transferring the program to other locales or jurisdictions; and (4) determine the extent to which the program results can be generalized. In terms of generalization, it is important not only to recommend that the program be promulgated, but also to define the limits of such a recommendation. Fourth, the first three systemic perspectives can be regarded as *program oriented* in focus as compared to the fourth perspective, which assesses the program results from a broader *policy oriented* perspective. In addition to assessing the policy implications, it is important to address other feasible and beneficial alternatives to the program. The alternatives could range from slight improvements to the existing program to recommendations for new and different programs.

The second part of the measures framework concerns the linkages among the various evaluation measures. A model of these linkages should contain the hypothesized relationships, including cause-and-effect relationships, among the measures. The model should help in identifying plausible test and rival hypotheses, as well as in identifying critical points of measurement and analysis. In practice, the model could simply reflect a systematic thought process undertaken by the evaluator, or it could be explicitly expressed in terms of a table, a block diagram, a flow diagram, or a matrix.

In conclusion, concise and measurable measures can mitigate the measures-related threats to validity. Additionally, the linkage model can help to avert some of the other threats to validity.

Measurement Methods. The list of issues and elements that constitute the measurement methods component includes measurement time frame (i.e., evaluation period, measurement points, and measurement durations), measurement scales (i.e., nominal, ordinal, interval, and ratio), measurement instruments (i.e., questionnaires, data collection forms, data collection algorithms, and electromechanical devices), measurement procedures (i.e., administered questionnaires, implemented data collection instruments, telephone interviews, face-to-face interviews, and observations), measurement samples (i.e., target population, sample sizes, sampling technique, and sample representativeness), measurement quality (i.e., reliability, validity, accuracy, and precision), and measurement steps (i.e., data collection, data privacy, data codification, and data verification).

Clearly, each of the above indicated measurement elements has been the subject matter of one or more theses, journal articles, and/or books. For example, data sampling, a technique for increasing the efficiency of data gathering by the identification of a smaller sample that is *representative* of the larger target data set, remains a continuing hot research

area in statistics. The dilemma in sampling is that the larger the sample, the greater the likelihood of representativeness *but,* likewise, the greater the cost of data collection.

Measurement methods that could mitigate or control for threats to validity include a multimeasurement focus, a long evaluation period (which, while controlling for regression artifacts, might aggravate the other threats to internal validity), large sample sizes, random sampling, pretest measurement, and, of course, techniques that enhance the reliability, validity, accuracy, and precision of the measurements. Further, judicious measurement methods can control for the test-setting sensitivity threat to external validity, while practical measurement methods that take into account the political and economic constraints can control for the conduct conclusion threats to validity.

Analytic Techniques. Analytic techniques are employed in evaluation or analysis for a number of reasons: to conduct statistical tests of significance; to combine, relate, or derive measures; to assist in the evaluation conduct (e.g., sample size analysis, Bayesian decision models); to provide data adjustments for nonequivalent test and control groups; and to model test and/or control situations.

Next to randomization (which is usually not implementable), perhaps the single most important evaluation design element (i.e., the one that can best mitigate or control for the various threats to validity) is, as alluded to above, modeling. Unfortunately, most evaluation efforts to date have made minimal use of this simple but yet powerful tool. Larson (8), for example, developed some simple structural models to show that the *integrity* of the Kansas City Preventive Patrol Experiment was not upheld during the course of the experiment—thus casting doubt on the validity of the resultant findings. As another example, Tien (9) employed a linear statistical model to characterize a retrospective "split area" research design or selection scheme, which was then used to evaluate the program's impact. The next section further underscores the importance of evaluation modeling.

EVALUATION MODELING

An important area in which evaluation modeling has played a critical role is criminal recidivism, which can be defined as the reversion of a person to criminal behavior after he or she has been convicted of a prior offense, sentenced, and (presumably) corrected. In particular, there have been many evaluations of correctional programs to determine if they work—more specifically, do they reduce the rate of recidivism? Maltz and Pollack (10), for example, show how a population of youths, whose delinquent activity is represented by a stationary stochastic process, can be selected (using reasonable selection rules) to form a cohort that has an inflated rate of delinquent activity before selection. When the activity rate returns to its uninflated rate after the youths are released from the program, an apparent reduction results. Based on this analysis, they conclude that the reductions noted in delinquent activity may be largely due to the way delinquents are *selected* for correction rather than to the effect of the programs. Thus, they modeled the impact of the regression artifact threat to internal validity.

Ellerman, Sullo, and Tien (11), on the other hand, offer an alternative approach to modeling recidivism by first determining empirical estimates of quantile residual life (QRL) functions, which highlight the properties of the data and serve as an exploratory aid to screening parametric mixture models. The QRL function can be defined as follows. Let T be a random variable (rv) which represents time-to-recidivism and F be its distribution function (df); thus,

$$F(t) = P(T \leq t) \tag{1}$$

Assume that F is absolutely continuous on its interval of support so that its derivative, denoted by f, is the probability density function (pdf) of T. The *reliability* or *survivorship function* (sf), denoted by $\overline{F}$, is

$$\overline{F}(t) = P(T > t) = 1 - F(t) \tag{2}$$

$\overline{F}(t)$ is the probability that an ex-prisoner will not recidivate before time t. Let T_x be the time remaining to recidivism given that an ex-prisoner has not yet recidivated at time x or, alternatively, the *residual life at time x*, that is,

$$T_x \equiv (T - x)|\{T > x\} \tag{3}$$

T_x is a conditional rv with sf $\overline{F}_x$ defined by

$$\begin{aligned} \overline{F}_x(t) &\equiv P(T - x > t | T > x) \\ &\equiv \overline{F}(t + x)/\overline{F}(x) \end{aligned} \tag{4}$$

The df of T_x is then given by $F_x = 1 - \overline{F}_x$. For any m in (0, 1), let

$$Q_m(x) = F_x^{-1}(m) \tag{5}$$

where $F_x^{-1}(\cdot)$ denotes the inverse function of F_x. Since the df F is assumed to be absolutely continuous, its inverse exists uniquely and hence so does that of the residual life df F_x. The function $Q_m(\cdot)$ is called the *m-quantile residual life* (QRL) *function;* $Q_m(x)$ is the m-quantile of the df F_x. For example, while $F^{-1}(0.5)$ is the median of the underlying distribution F, $Q_{0.5}(x)$ is the median of the residual life distribution F_x. Simple distinctions such as whether $Q_m(x)$ is increasing or decreasing in x for any m is tantamount to a statement concerning recidivism dynamics.

It follows from Eqs. (4) and (5) that in terms of the unconditional df F,

$$Q_m(x) = F^{-1}[(1 - m)F(x) + m] - x \quad \text{for } x \geq 0 \tag{6}$$

If $Q_m \equiv Q_m(0) = F^{-1}(m)$ denotes the m-quantile of the original distribution, then Eq. (6) is equivalent to

$$Q_m(x) = Q_{m'} - x \tag{7}$$

where

$$m' = (1 - m)F(x) + m \tag{8}$$

For any distribution for which the inverse function F^{-1} exists in closed form, Eq. (6) will yield closed-form expressions for the QRL functions. It can be shown, under some mild condi-

tions, that the mean of the distribution is infinite if for any m

$$\lim_{x\to\infty} dQ_m(x)/dx > m/(1-m) \tag{9}$$

A common notion in the stochastic recidivism literature is, either by explicit assumption or by inference from the fitted distributions, that recidivism rates decline over time. Applied to social systems or processes, this phenomenon has been called *inertia.* The concept of inertia is that for $y \geq x$, $T_y \geq T_x$ in *some sense.* The strongest such sense is that of a *decreasing hazard rate* (DHR). The most fundamental way of stating this condition is

$$\overline{F}_x(t) = P(T - x > t | T > x) \uparrow in\, x \quad \text{for all } t > 0 \tag{10}$$

where $\uparrow$ means nondecreasing. The property in Eq. (10) is equivalent to the statement, for $y \geq x$, T_y is stochastically greater than T_x. It is also equivalent to the nonincreasingness of the hazard rate $\lambda(t) = f(t)/\overline{F}(t)$. The DHR property of Eq. (10) states that the longer an ex-prisoner remains free, the more probable is it that he remains free for an additional time t. It can be shown that Eq. (10) holds if, and only if, $Q_m(x) \uparrow$ in x for all m, that is, the DHR property is equivalent to the nondecreasingness of every QRL function. Thus, it is conceivable that $Q_m(x) \uparrow$ in x for some m, say the median, so that inertia would be present with respect to the median residual life even when the underlying distribution is not DHR. The DHR property has a dual, namely, the *increasing hazard rate* (IHR) property defined by the substitution of $\downarrow$ for $\uparrow$ in Eq. (10). It should be noted that while the terminology and applications contained in this section pertain to the criminal justice area, the proposed QRL approach can be applied in other contexts as well, for example, nursing home length of stay, disease latency and survivability, and reliability engineering.

Empirical estimates of quantile residual life functions can be employed not only to obtain properties of recidivism, but also to help screen parametric mixture models. In this manner, the Burr model is demonstrated to be an appropriate model for characterizing recidivism. The Burr is actually a mixture of Weibull distributions; its sf is

$$\overline{F}(t) = (1 + \beta t^{\rho})^{-\alpha},\ t > 0 \tag{11}$$

Where $\alpha, \beta, \rho > 0$. The hazard rate of the Burr is

$$\lambda(t) = \alpha\beta\rho t^{\rho-1}/(1 + \beta t^{\rho}),\ t > 0 \tag{12}$$

which is strictly decreasing for $\rho \leq 1$ (as expected because then the Weibulls being mixed are DHR), while for $\rho > 1$, that is, for a mixture of IHR Weibulls, $\lambda(t)$ is increasing on $[0, x_r]$, where

$$x_r = [\beta^{-1}(\rho - 1)]^{1/\rho} \tag{13}$$

and then decreases on (x_r, ∞). Thus, while mixtures of nonincreasing hazard rate distributions are strictly DHR, mixtures of IHR distributions are not necessarily IHR.

As applied to criminal recidivism, then, the Burr model suggests that although the observed declining recidivism rate can be explained by population heterogeneity, individual recidivism rates may in fact be increasing. This understanding can have a significant impact on public policy. For example, the observation that there is an initial high recidivism rate among cohorts of prison releasees has led some criminologists to contend that adjustment problems (i.e., "postrelease trauma") during a "critical period" soon after release result in intensified criminal activity. This observation has resulted in various correctional programs, such as halfway houses, intensified parole supervision, and prison furloughs, designed to alleviate postrelease stress and minimize recidivism. But these programs may be based on a misinterpretation of recidivism data. There may very well be a "critical period" after release; however, the observed high recidivism rate during the critical period, followed by what seems to be a declining rate, may be an artifact of population heterogeneity and, if so, should not be associated with individual patterns of recidivism. Thus, basing postrelease supervision programs on inferences made from aggregate data—when such inferences concern individual behavior—is risky business.

In sum, evaluation modeling is critical to any system or program evaluation. In many situations, as is the case above for criminal recidivism, it provides for a "control" framework within which the system or program performance is analyzed or understood.

OBSERVATION

In the continued development and operation of a system or program, it is obvious that broad evaluation efforts cannot be continuously carried out; indeed, such efforts should only be undertaken on an intermittent—and as needed—basis. The question then arises: What, if anything, should be done in between these system evaluations?

The answer can perhaps be found in the health field; in particular, in the way a physician conducts a physical examination. At the start of the examination, the doctor checks a basic set of indicators (e.g., blood pressure, temperature, heart rate). If any of these primary indicators signals a potential problem, measurements of other indicators that dig deeper into the body's systems are taken (e.g., blood test, x-ray, CT scan). If any of these secondary indicators suggests a problem, then other tertiary indicators (e.g., colonoscopy, biopsy) may be ascertained. As the doctor digs deeper and deeper, the root cause of the problem or symptom is discovered and appropriate actions are taken to correct the problem. In other words, a layered approach is taken toward monitoring, diagnosing, and improving a person's physical health. Similarly, for example, in assessing the "health" or performance of a system, a layered approach could be employed, starting with broad, easy-to-obtain measures and continuing, if necessary, with more focused measures. In fact, as suggested in Fig. 1, three layers of measures, primary, secondary, and tertiary, would probably be sufficient. This approach should also include a method for combining at least the primary measures into, say, a system performance index (SPI) that could be used to help assess the system status on an ongoing, continuous basis, just as the Dow Jones Industrial Average serves to gauge stock market performance on a continuous basis. A system with a low SPI would need to acquire secondary and/or tertiary measures in order to identify appropriate strategies for improving its SPI.

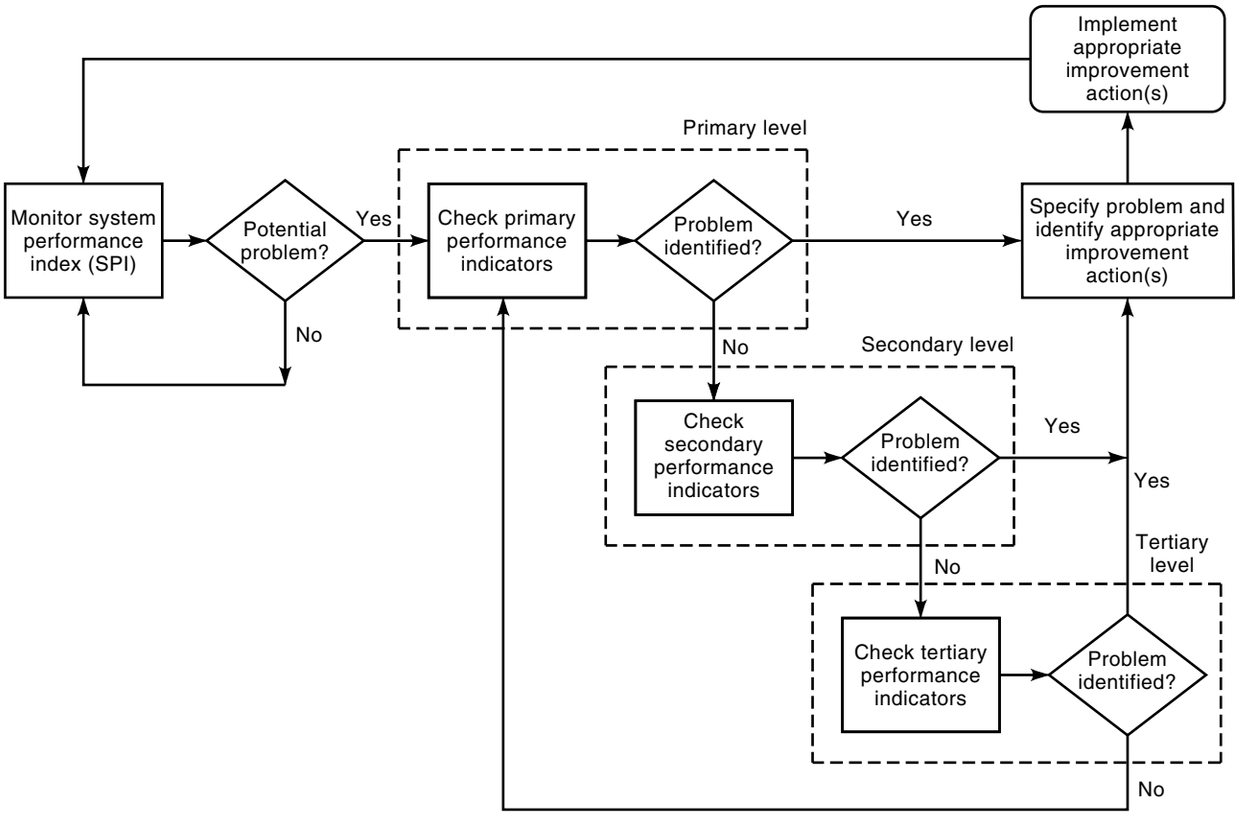

Figure 1. A continuous layered approach to the monitoring, diagnosis, and improvement of systems.

The continuous layered approach to the monitoring, diagnosis and improvement of systems (CLAMDIS) can only be promulgated if the initial system evaluation yields a pertinent set of primary, secondary, and teritary indicators and demonstrates how the primary indicators can be combined into an overall SPI. An effective system evaluation must identify such pertinent indicators; otherwise, the evaluation would be, at best, a one-time assessment of the system's performance. In this regard, CLAMDIS would be complementary to the evaluation approach presented herein.

BIBLIOGRAPHY

1. E. Chelimsky and W. R. Shadish, Jr. (eds.), *Evaluation for the 21st Century,* Thousand Oaks, CA: Sage Publications, Inc., 1997.
2. J. M. Tien, Program evaluation: A systems and model-based approach, in A. P. Sage (ed.), *Concise Encyclopedia of Information Processing in Systems and Organizations,* New York: Pergamon, 1990.
3. J. M. Tien, Toward a systematic approach to program evaluation design, *IEEE Trans. Syst. Man. Cybern.,* **9**: 494–515, 1979.
4. K. W. Colton, M. L. Brandeau, and J. M. Tien, *A National Assessment of Command, Control, and Communications Systems,* Washington, DC: National Institute of Justice, 1982.
5. J. M. Tien and M. F. Cahn, Commercial security field test program: A systematic evaluation of the impact of security surveys, in D. P. Rosenbaum (ed.), *Preventing Crime in Residential and Commercial Areas,* Beverly Hills, CA: Sage Publications, 1986.
6. J. M. Tien and T. F. Rich, *Early Experiences with Criminal History Records Improvement,* Washington, DC: Bureau of Justice Assistance, 1997.
7. D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research,* Chicago: Rand McNally, 1966.
8. R. C. Larson, What happened to patrol operations in Kansas City? A review of the Kansas City preventive patrol experiment, *J. Crim. Just.,* **3**: 267–297, 1975.
9. J. M. Tien, Evaluation design: Systems and models approach, in M. G. Singh (ed.), *Systems and Control Encyclopedia,* New York: Pergamon Press, 1988.
10. M. D. Maltz and S. M. Pollock, Artificial inflation of a delinquency rate by a selection artifact, *Op. Res.,* **28** (3): 547–559, 1980.
11. R. Ellerman, P. Sullo, and J. M. Tien, An alternative approach to modeling recidivism using quantile residual life functions, *Op. Res.,* **40** (3): 485–504, 1992.

JAMES M. TIEN
Rensselaer Polytechnic Institute